DOCUMENT RESUME

ED 455 296	TM 033 153
AUTHOR	Leimu, Kimmo, Ed.; Linnakyla, Pirjo, Ed.; Valijarvi, Jouni, Ed.
TITLE	Merging National and International Interests in Educational System Evaluation. Proceedings of the Conference (Jyvaskyla, Finland, March 19th and 20th, 1998).
INSTITUTION	Jyvaskyla Univ. (Finland). Inst. for Educational Research.
ISBN	ISBN-951-39-0915-8
PUB DATE	2001-00-00
NOTE	150p.
AVAILABLE FROM	Institute for Educational Research, Customer Services, University of Jyvaskyla, P.O. Box 35, FIN-40351 Jyvaskyla, Finland. Tel: +358-14-260-3220; Fax: +358-14-260-3241; e-mail: teairmajyu.fi; Web site: http://www.jyu.fi/ktl.
PUB TYPE	Collected Works - Proceedings (021)
EDRS PRICE	MF01/PC06 Plus Postage.
DESCRIPTORS	*Educational Quality; Elementary Secondary Education; *Evaluation Methods; Foreign Countries; Higher Education; International Education; *International Studies; *Systems Analysis
IDENTIFIERS	Educational Indicators

ABSTRACT

Papers from this conference focus on acquiring and using empirical information as a basis for monitoring and studying education with the special ambition of making such information both meaningful and powerful and the use of such information dynamic. The papers are: (1) "The Way to a Strategic View on Evaluation" (Kimmo Leimu); (2) "Strategic Arenas of Influence in Pursuing Quality in Education: Some Conceptual and General Issues" (Ulf P. Lundgren); (3) "The Potential and Challenges of International Comparative Studies of Educational Achievement" (Tjeerd Plomp); (4) "Educational Indicators" (Eugene Owen); (5) "National Viewpoints on International Evaluation and Research" (Erkki Kangasniemi); (6) "The National Intertwined with the International" (Pirjo Linnakyla); (7) "Staking Claims for Quality in System Evaluation in Germany" (Rainer Lehmann); (8) "Challenges to a National Strategy of Evaluation: Visions and Expectations" (Vilho Hirvi); (9) "How To Use Evaluation Findings" (Pentti Takala); (10) "Pillars of National Evaluations: Reconciling Research and Policy Interests in Evaluation Programs in Finland" (Pentti Yrjola); (11) "Research in the Context of National Assessment" (Jouni Valijarvi); and (12) "Conclusions, Challenges, and Visions" (Kimmo Leimu). Each paper contains references. (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 455 296

PERMISSION TO REPRODUCE AND
DISSEMINATE THIS MATERIAL HAS
BEEN GRANTED BY

R. Pitkänen

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)

1

Edited by
Kimmo Leimu
Pirjo Linnakylä
Jouni Välijärvi

Merging NATIONAL and INTERNATIONAL interests

in educational system evaluation

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

BEST COPY AVAILABLE

TM033153

INSTITUTE FOR
EDUCATIONAL RESEARCH
UNIVERSITY OF JYVÄSKYLÄ

MERGING NATIONAL AND INTERNATIONAL INTERESTS IN EDUCATIONAL SYSTEM EVALUATION

MERGING NATIONAL AND INTERNATIONAL INTERESTS IN EDUCATIONAL SYSTEM EVALUATION

Proceedings of the Conference
held at the University of Jyväskylä, Finland
on March 19th and 20th, 1998

Edited by
Kimmo Leimu, Pirjo Linnakylä & Jouni Välijärvi

INSTITUTE FOR EDUCATIONAL RESEARCH
UNIVERSITY OF JYVÄSKYLÄ

Advisory Board of the Publication Series of the
Institute for Educational Research
Jouni Välijärvi
Pirjo Linnakylä
Viking Brunell
Päivi Häkkinen
Päivi Tynjälä
Jouni Sojakka

THIS PUBLICATION
CAN BE OBTAINED FROM:
Institute for Educational Research
Customer services
University of Jyväskylä
P.O. Bpx 35
FIN-40351 Jyväskylä, Finland
Phone +358 14 260 3220
Fax +358 14 260 3241
E-mail: teairamajyu.fi
www.jyu.fi/ktl

© Authors and Institute for Educational Research

Editors: Kimmo Leimu, Pirjo Linnakylä & Jouni Välijärvi
Editorial assistant: Inga Arffman

Cover: Martti Minkkinen
Layout: Uuve Södor

ISBN 951-39-0915-8

Printed by University Printing House and ER-paino Ky (covers)
Jyväskylä 2001

Contents

Preface

THEME 1:
AN INTERNATIONAL CONTEXT OF SYSTEM EVALUATION

The Way to a Strategic View on Evaluation
Kimmo Leimu 7

Strategic Arenas of Influence in Pursuing Quality in Education: Some Conceptual and General Issues
Ulf P. Lundgren 15

The Potential and Challenges of International Comparative Studies of Educational Achievement
Tjeerd Plomp 23

Educational Indicators
Eugene Owen 41

THEME 2:
MERGING NATIONAL AND INTERNATIONAL CONCERNS

National Viewpoints on International Evaluation and Research
Erkki Kangasniemi 53

The National Intertwined With the International
Pirjo Linnakylä 63

Staking Claims for Quality in System Evaluation in Germany
Rainer Lehmann 77

THEME 3:
A NATIONAL STRATEGY OF EVALUATION

Challenges to a National Strategy of Evaluation:
Visions and Expectations
Vilho Hirvi 93

How to Use Evaluation Findings
Pentti Takala 99

Pillars of National Evaluations: Reconciling Research and Policy
Interests in Evaluation Programs in Finland
Pentti Yrjölä 107

THEME 4:
MERGING NATIONAL ASSESSMENT AND RESEARCH POLICIES

Research in the Context of National Assessment
Jouni Välijärvi 113

Conclusions, Challenges and Visions
Kimmo Leimu 123

Preface

With the advent of government-sponsored monitoring activities pertaining to education, system level statistics, assessment results and comparative studies are getting increasing attention from the general public in several countries. While a great number of national units have been involved in international co-operative studies for decades already, notably in the form of IEA research, such sources of information are nowadays becoming increasingly widespread and visible, due to the increasingly recognized official nature of assessment activities. Apart from regional (e.g. European, Asian) co-operative endeavors, the worldwide role of the OECD and its Indicators of National Education Systems (INES) program has lately become a significant source of information in the field of international studies. This development is supported by the OECD's widespread publications – such as *Education at a Glance*, or *Education Policy Analysis* – which are offering data on several interesting comparative issues, including the economics of education and the lifelong learning perspective. Other distinctive features of this "new" approach to international databases, in comparison with previous IEA research work, include its aspirations for increased regularity and up-to-dateness, which have been difficult to achieve in bona fide research projects relying on less stable non-governmental funding. These new developments not only offer regular sources of information, but also create a continuous need for fresh and reliable data on the state of education around the world.

This situation with a multiplicity of offerings can be welcomed. However, with widening horizons and ever higher aspirations, a number of inevitable concerns become evident, not least at the national level. There are signs of an increasing overlap and thus of a conflict between several simultaneous national and international endeavors involving either monitoring or research, each with its quality requirements which are becoming increasingly stringent. All of this will make heavy demands on time, money, and human resources, which may be regarded as the necessary prerequisites of quality. But the sheer volume of effort and products is not the only consideration. Another will be the nature and content of the entire evaluation approach – negotiating the Scylla and Charybdis of basic differences in interests between monitoring and research. What may be won by freshness, general appeal and tight schedules necessitated by a few regularly produced indicators, may be compromised by their exorbitant costs and lack of depth, while leaving most of the further questions unanswered. Among other issues to be resolved is the comprehensive *modus operandi,* a national strategy for monitoring, assessing and understanding the nature, process and progress of education. This will require some notion of a national evaluation strategy, which should somehow be harmonized with international co-operative efforts.

While most of the time in negotiating the bigger and smaller issues of international comparative work is spent on technical and communication issues, the need to consider the above fundamental watersheds is evident. Particularly in the context of ongoing massive programs involving not only international student assessment on a regular basis (OECD/PISA), but other multilevel actors and factors on the national educational scene as well, the need for open discussion has emerged. Although much of the development work is taking place in terms of international co-operation, such negotiations should not be limited to cross-cultural fora only, because the basic needs and aspirations are an equally serious concern at the national and local levels as well. It was therefore decided that opportunities for a broadly conceived discussion should be offered. This publication aims at opening doors to achieve these ends.

Kimmo Leimu Pirjo Linnakylä Jouni Välijärvi

THEME 1

AN INTERNATIONAL CONTEXT OF SYSTEM EVALUATION

Kimmo Leimu
Institute for Educational Research
University of Jyväskylä
Finland

The Way to a Strategic View on Evaluation

General Background

This publication is a child of love of one rather specialized sector in education – that of *acquiring and using empirical information as a basis for monitoring and studying education with the special ambition of making such information both meaningful and powerful, and its use dynamic.* This pursuit has a distinguished past, an intriguing present, and a challenging future. In particular, the present publication has been inspired – and even *necessitated* – by two major approaches to these ideals:

- The work of the IEA, the International Association for the Evaluation of Educational Achievement on comparative educational research, which started in the 1950s and is going on. It may be appropriate to remind that Finland was one of the IEA's founding members and that the Institute for Educational Research (IER) at the University of Jyväskylä has been its national member institute all along.
- The other and more recent one is the OECD/CERI project INES for developing and acquiring Indicators of National Education Systems, especially its strategy Network A, which aims at collecting student learning outcomes data on a regular basis. Again, the IER was assigned responsibility for the national conduct of the INES data strategy in Finland.

It is therefore evident that we can start discussing educational system assessment from experience, and not from its very beginning. This also implies that it *is no more a completely open field,* but one in which certain premises and traditions have already been established. At the same time, several issues and principles are still waiting for further solutions.

For reasons related to the basic interests and the general nature of the work, *the focus will now be on system level studies and assessments.* This is not to deny the importance of work at every level of education, but simply to provide an opportunity to concentrate on a field of growing interest in our global village, in order to give it a fair chance and the necessary thought.

The Rationale and Particular Intentions

Everywhere in the world, educational investments are made expecting positive returns on efforts which are meant to have long-term effects on the nation's daily work and economy, the quality of life of the citizens, and on the entire culture. While every individual's emerging personality and whole life are fundamentally at stake, from society's point of view the educational effort will inevitably absorb considerable proportions of any nation's capacity for maintenance and regeneration – even for survival. This can be traced back to the now self-evident demand for ever higher levels of formal and non-formal *education for all,* which makes the expectations both intensive and extensive. Both symbolically and economically, *education is dear* to most nations.

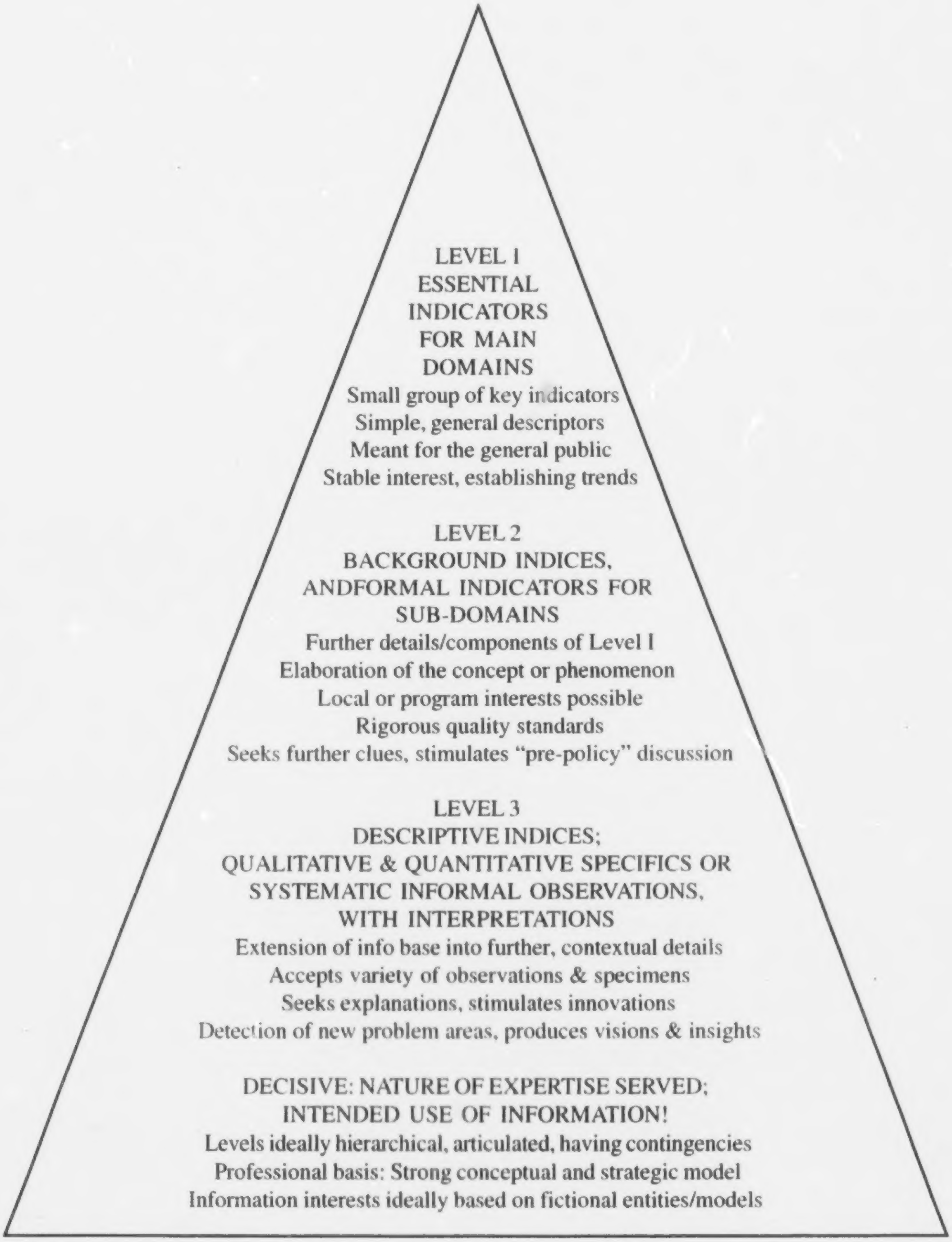
Since formal – and especially basic – education is concerned with entire age groups and since much of it is based on public funding, *knowledge of the effort and its outcomes* have become *a legitimate concern* of not only those who are accountable for decisions and actions in education, but increasingly of anyone involved in and affected by the process.

A number of explicit *criteria for* the desired *quality and effects of education* are presented in public speeches and writings, legal documents and more spontaneously among the users. These include several important e's: equality, excellence, efficiency, effectiveness, economy, and empowerment – while all of this should also be enlightening, enjoyable and even entertaining, making use of and leading to considerable expertise, without being extremist. Despite constant concerns, much of the *actual attainment* of such noble aims tends to remain with shallow evidence, to say nothing of testable proof. While it would be foolish for anyone to walk blindfolded in difficult terrain by hearsay only, we often tend to take this risk in one of our major human and material investments – *preparation for life*. Lives of individuals and of entire societies. Some of the *reasons* can perhaps be directed at us educators – policy makers, practitioners, and researchers – for not seeing clearly enough and insisting on some of the evident needs of this important national effort, when perceived as a complex, multilevel organization.

There is no denying that *the role of assessment and evaluation in education has long been recognized.* However, the main concerns have traditionally been at the fundamental level of teaching and learning, where the focus is on individuals and their credentials. Accordingly, student assessment and examinations have long dominated the evaluation scene, not least as instruments of selection and rejection. This has become its "historical" encumbrance, the effects of which are deeply felt in almost any evaluation effort. Almost too easily, evaluation discourse still tends to slip into "certifying" concerns for good or bad, pass or fail, or simplistic judgements of quality, not to mention fears of labeling, and mysterious threats of sanctions, especially as regards clearly identified persons, institutions, or countries, without ever asking for deeper diagnosis or elaboration. These are indications of the often prevailing conception of evaluation as a tool of authority and power, rather than as one of professional diagnosis and well-reasoned support. (Cf. Norris, 1995; Pettersson & Wallin, 1995; Vedung, 1995.) All of this calls for an adequate *evaluation culture.*

Interest in the *evaluation of programs or entire educational systems* is of much more recent origin – by and large a post-World-War-II phenomenon. Although one might have to admit that this is among the youngest fields of specialized expertise in any nation's education endeavor, (most of us would, in fact, represent only the second or third generation of professional evaluators or researchers), it is, however, no more virgin territory either in terms of practice or preaching. As prime examples we can take the two major international co-operative efforts represented here today: *IEA research* on student achievement, which had its beginnings in the 1950s, and the *OECD educational indicators project,* which first raised its head in the early 1970s and had a more determined restart in the late 1980s. Also *at the national level,* several countries (for instance, the USA, Australia, France, Italy and Sweden) have for some time now had a tradition of national assessments with comprehensive aims and ambitions. Problems and principles of evaluation have thus

Figure 1. The information pyramid idea by Bryk and Hermanson (1994).



LEVEL 1
ESSENTIAL
INDICATORS
FOR MAIN
DOMAINS
Small group of key indicators
Simple, general descriptors
Meant for the general public
Stable interest, establishing trends

LEVEL 2
BACKGROUND INDICES,
ANDFORMAL INDICATORS FOR
SUB-DOMAINS
Further details/components of Level I
Elaboration of the concept or phenomenon
Local or program interests possible
Rigorous quality standards
Seeks further clues, stimulates "pre-policy" discussion

LEVEL 3
DESCRIPTIVE INDICES;
QUALITATIVE & QUANTITATIVE SPECIFICS OR
SYSTEMATIC INFORMAL OBSERVATIONS,
WITH INTERPRETATIONS
Extension of info base into further, contextual details
Accepts variety of observations & specimens
Seeks explanations, stimulates innovations
Detection of new problem areas, produces visions & insights

DECISIVE: NATURE OF EXPERTISE SERVED;
INTENDED USE OF INFORMATION!
Levels ideally hierarchical, articulated, having contingencies
Professional basis: Strong conceptual and strategic model
Information interests ideally based on fictional entities/models

BEST COPY AVAILABLE

been discussed for a considerable time at every level – the institutional, municipal, provincial, and system level.

However, recognition of the need for an organized, multilevel approach to evaluation or insight into the complementary functions of different sources of information is of a more recent vintage. While it is evident that issues as complex as those concerning education can hardly be described in simple terms, the mere acceptance of a serious use of feedback information necessitates a particular environment: a *climate open to self-criticism and change,* with a view to renewal and development, which is built upon certain *rationalism* and *democratic principles* of transparency, openness and valuation of the common good, while respecting local and individual efforts. Taken further, this welcomes a *culture of evaluation* in which a wide variety of interests and paradigms are accepted and nurtured, including both systematic and representative data ideals and less formalistic participatory approaches.

All of this may be deemed necessary if one is to obtain a rich enough picture of education, perhaps resulting in what Bryk and Hermanson (1994) have called *the information pyramid* (Figure 1). It can be well understood that databases *cannot* be acquired *ad hoc,* without careful planning and preparation, or *alone,* by any single individual, institution, or – I am tempted to say today – even by a single nation.

Evaluation Practices

There are several ways of looking at evaluation practices and their developments. One may follow the footsteps of Guba and Lincoln (1989) and review the various *stages in the history of educational evaluation,* as they are sometimes presented (e.g. Guba & Lincoln, 1989; Konttinen, 1995; Rombach & Sahlin-Andersson, 1995).

1. *Evaluation as measurement (of separate facts):* technical and straightforward orientation without much problematization of aims or procedures. Interpretation left for the client to make, not among evaluator responsibilities.
2. *Evaluation as monitoring well-defined pursuits, goals and objectives, and as support to decision-making.* The entire society was considered a field of experimentation or the world an educational laboratory. Results were related to the set aims, which were left unquestioned. The results themselves were consid-

ered valid, objective, and reliable, their interpretations were seen as impartial judgements, rather directly useful in decision-making.

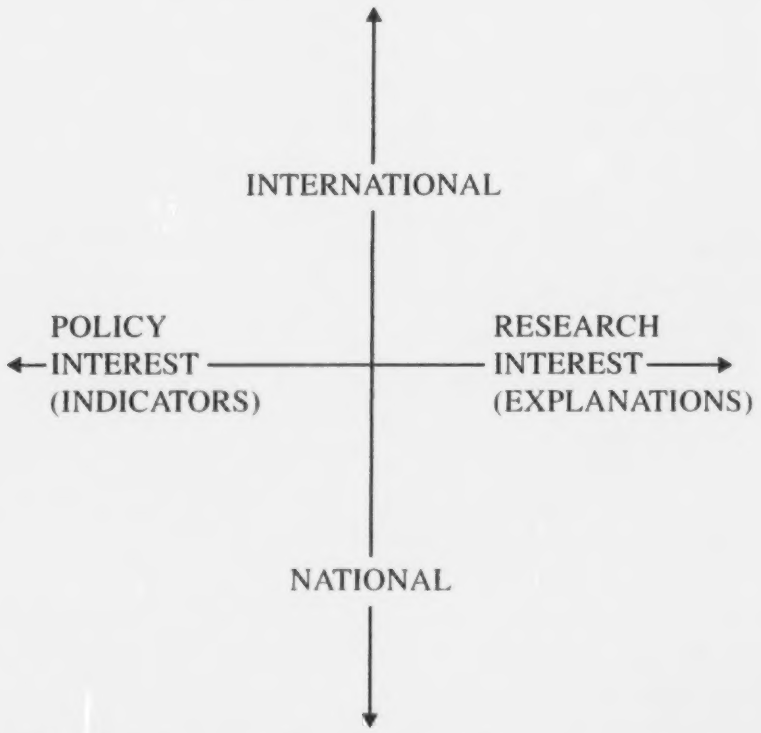
3. *Increasing demands for impartiality and externality in evaluation* became evident, and an *authentic research element* was seen as the main proof of its quality. Whereas scientific concepts and problems were introduced, *evaluations still remained "above* the situation". Later on, *different approaches to and perspectives on the same problem were recognized* – even demanded. So much so that initial assumptions were taken into account and a variety of viewpoints could be presented as outcomes. *Theory-based evaluation was called for* and the active contribution of various parties of interest was recognized.
4. *Responsive constructivistic evaluation* means an ever-increasing pursuit to handle complexity. It is expected that *evaluation becomes a dialogue* among different interest groups – in other words, *evaluation is seen as a process, or an arena for exchanging viewpoints* and experiences. Among the prerequisites, the models of steering in the particular system in which the evaluation is taking place become important.
5. One may wish to envision *one further step* – or at least an elaboration – among these views on evaluation, emphasizing the influences and *role of societal changes:* the growth of internationalism (global integration) and new technologies, but also the deregulation and enhancement of ownership through the principle of subsidiarity and thus more distributed professionalism. At the system level, this is paving the way for *a strategic view on evaluation.*

In this view, system evaluation would *not* be understood as a one-time, one-shot effort. *Instead,* it would be seen as a more comprehensive *offensive* which has both immediate and more long-term or fundamental aims, as well as a multilevel participation structure. Evaluation premises and their outcomes would here be *put in a broader context,* and as a consequence, a higher degree of planning and conceptualization would be necessary.

The vision itself would be to incorporate information related to several different, but conceivably associated problem areas at several levels of operation, and allowing – but also requiring – different paradigms (both formal and informal). The strategy would necessitate a rather comprehensive model of ends and means, effects and influences. All of this would hardly be possible at one go, but instead requires *a stepwise or multilevel strategic approach,* where one information need would arise from another and through an *elaborative knowledge-hunting* endeavor would lead to a gradually completed mosaic. – *The ultimate aim here would be, not* to develop a better *evaluation* system, but to develop a better functioning *educational system,* schools which see their mission more clearly, with teachers and students who may attain higher quality learning. – While such a strategic approach would demand a variety of expertise and data, *it could never be accomplished alone.* Whether using self-assessments or external evaluations, *the necessary partnerships* should be developed as tools for this.

Should this final vision sound too utopian, I am inclined to say *we are well on our way to it* already.

Figure 2. The principal problem dimensions.


INTERNATIONAL
POLICY INTEREST (INDICATORS)
RESEARCH INTEREST (EXPLANATIONS)
NATIONAL

As indicated before, *two principal problem dimensions* at this point of developing international assessment activities will warrant in-depth discussion: one concerned with *research vs. policy indicator interests* in the system evaluation arena, the other with *international vs. national interests* in conducting large-scale evaluation studies (see Figure 2). Both of these problem areas have to do with and reflect national strategies and implementation arrangements because any co-operative international activity will necessarily be based on national resources and work contributions. It will therefore be necessary to start thinking about priorities and strategies, but also *capacity-building* – the recruitment and training of the necessary human resources – and eventually the problems of organization and funding.

References

Bryk, A., & Hermanson, K. (1994). Observations on the structure, interpretation and use of the education indicator system. In OECD/CERI, *Making education count: Developing and using international indicators* (pp. 37-53). Paris: Author.

Guba, E. G., & Lincoln, Y. S. (1989). *Fourth generation evaluation*. Newbury Park, CA: Sage Publications.

Konttinen, R. (1995). Arvostelusta näyttöön – Koulutuksen arvioinnin kehityspiirteitä Suomessa [Changing interests in evaluation – Evaluation research in Finland]. In S. Takala (Ed.), *Arviointi ja koulutuksen laadun kehittäminen* [Evaluation and the enhancement of the quality of education] (pp. 9-22). Jyväskylä: University of Jyväskylä, Institute for Educational Research.

Norris, N. (1995). Koulutusohjelmien arviointi [Evaluation of educational programmes]. In S. Takala (Ed.), *Arviointi ja koulutuksen laadun kehittäminen* [Evaluation and the enhancement of the quality of education] (pp. 33-38). Jyväskylä: University of Jyväskylä, Institute for Educational Research.

Pettersson, S., & Wallin, E. (1995). Utvärderingsmakt [The power of evaluation]. In B. Rombach, & K. Sahlin-Andersson (Eds.), *Från sannigssökande till styrmedel. Moderna utvärderingar i offentlig sektor* [From seeking after the truth to steering methods. Modern evaluations in the public sector](pp. 93-108). Stockholm: Nerenius & Santerus Förlag.

Rombach, B., & Sahlin-Andersson, K. (1995). På tal om utvärdering [Speaking of evaluation]. In B. Rombach, & K. Sahlin-Andersson (Eds.), *Från sannigssökande till styrmedel. Moderna utvärderingar i offentlig sektor* [From seeking after the truth to steering methods. Modern evaluations in the public sector] (pp. 9-24). Stockholm: Nerenius & Santerus Förlag.

Vedung, E. (1995). Utvärdering och de sex användningarna [Evaluation and its six uses]. In B. Rombach, & K. Sahlin-Andersson (Eds.), *Från sannigssökande till styrmedel. Moderna utvärderingar i offentlig sektor* [From seeking after the truth to steering methods. Modern evaluations in the public sector] (pp. 25-51). Stockholm: Nerenius & Santerus Förlag.

Ulf P. Lundgren
National Agency for Education
Sweden

Strategic Arenas of Influence in Pursuing Quality in Education: Some Conceptual and General Issues

In my presentation I shall describe three conceptions of quality. One in which quality is defined as value judgements and is thus embedded in general ideas about the meaning and purpose of education, i.e. the very way we think of education and legitimate education and schooling. The other will deal with quality from the viewpoint of meeting given standards. The third concept is quality as a relation between the subject and the object and thus something that develops through enlightening discourse. Finally, I shall point at the use of international indicators as a consequence of how these three concepts are defined.

The Concept of Quality

The word quality is frequently used in everyday language. It is easy to use but hard to define. In *The Oxford Guide to the English Language* (1998), quality is defined as the "degree or level of excellence; [a] characteristic, something that is special in a person or thing". In industrial production, the term quality has come to mean "conformance to requirements". Thus, quality used according to that definition is something that meets specific standards. This definition is often embedded in the use of terms such as quality assurance.

The word quality itself stems from the Latin *qua litas*, which means the whole with its specific characteristics. The English word quality is not easy to translate into other languages, and when defined as conformance to requirements it becomes even more complicated. In Swedish the word *kvalitet* indicates a value judgement. This connotation is found in other Germanic languages as well.

Thus, the word quality and terms such as quality assurance and quality control are used ambiguously. On the one hand, quality is used as indicating a standard. The quality of McDonald's hamburgers means that they taste the same in New York and Stockholm. On the other hand, quality is a subjective judgement of specific characteristics. In this respect the quality of a hamburger is the taste of it. This ambiguity was considered early in Greek philosophy. Aristotle defined quality as something that convinces, i.e. as a relation between the object and the subject and thus something that belongs to both of them. Consequently, quality is seen as developing in discourse.

I shall use these three meanings of quality: quality as a subjective judgement; quality as standards; and quality as a relation between the subject and the object or as enlightening discourse.

Quality as a Judgement

Judgements about the quality of education are based on ideas about the meaning and the *raison d'être* of educational systems. As regards general education, the *raison d'être* can be formulated in terms of formation or by using the German word *Bildung*.

Behind every type of general education there are basic ideas about the kind of individual that is to be the fruit of the formation. In ancient Greece, this vision was expressed by the concept *Paidea*. General education had to be comprehensive and balance various areas of knowledge and skills and, in that way, create harmony between the mind and the body. The term *Paidea* was translated *humanitas* by Cicero and given another meaning in Roman education emphasizing rhetoric skills.

During the Middle Ages, fundamental ideas were molded, ideas that had an impact on educational thinking for centuries. The medieval religious *mysteqer* believed that the image of God was engraved in every human being. The German word *Bildung* stems from the notion "image" – *Bild*. As humans we were fallen angels; the meaning of life was man's struggle to preserve the image of God. During the Renaissance, the Greek concept of a well-balanced mind and body was restored in Italy. The ideal was *l'uomo universale*, the universal man to whom nothing was unfamiliar. As God had created everything, the logic was that knowing about the world was to know about God. It is in these thoughts that *Bildung* came to embrace the meaning of creation or formation, which is reflected in the English language.

During the 18th century, we can see two patterns of thinking taking form. In England and Scotland, a network of clubs was established and new types of magazines were founded. Bourgeois culture, formed along with these new arenas of discourse, developed a notion of the ideal gentleman, which in Anglo-Saxon countries became an essential part of the ideals of formation. In Germany, Kant published in 1784 in *Berliner Monatschrift* an article in which he linked the concepts of *Bildung* and liberation to each other. "Enlightenment is," he writes, "man's renouncing his own created incapacity." Kant's thoughts gave birth to the formation of the concept of modernity.

In modern thinking, knowledge became active and education was perceived from the angle of its pragmatic values: Education was thought to be a way of achieving tools to change and master the world. With these ideas about the *raison d'être* of education, the word *Bildung* was limited to cover mastery of the prescribed content of knowledge. In the German language, the new conception of education with its meaning and aims was denoted by the word *Ausbildung*. *Ausbildung* – in Swedish *utbildning* – has been the term for mass education motivated by its pragmatic consequences. We can see how these ideas were formed during the 19th century and clearly expressed during the 20th century by philosophers such as John Dewey. They were elaborated within the progressive education movement and came to have their general impact during post-war school reforms, at least in Anglo-Saxon and Nordic countries.

The purpose of this exegesis has been to focus on the context within which the concept of the quality of education has been given meaning. In educational thinking, where the ideals of education emphasize pragmatic values, the quality of education is judged in terms of how well education or schooling prepares students for citizenship and working life. Two centuries ago, the quality of education was judged in terms of how well it reproduced the "golden ages" of the Greek and Roman civilizations.

The Rise of Modernity

Each generation probably feels that the world is changing rapidly. A century ago (in 1899), Dewey published *The School and Society* based on lectures, in which he points to the rapidly changing society and the consequences these changes have for what kind of knowledge is needed, and he argues for the adjustment of education to this new society. In his lectures his arguments are that modern society becomes more and more invisible to children and the youth. These changes make new demands on formation and schooling in a modern world.

Looking in the mirror, there are striking parallels between the end of the 19th century and the end of the 20th century. Parallels in the sense that we are again facing a new world in which knowledge and the structure of knowledge and not least demands on knowledge are changing. And we are no more able now than we were a century ago to understand and cope with changes. Abstract conceptions are formed in the same way as they were a century ago even though the terms are different. We are not talking about a demanding and invisible world any more, instead we claim a world of knowledge – the knowledge or learning society. And we realize that we are entering a different society in which it is not easy, in so far as it is possible at all, to design and decide on the pragmatic values of a specific content of knowledge. Based on that insight, we formulate new ideas about and new terms for education, such as lifelong education. These new ideas about education and its *raison d'être* become a new construct telling of a new kind of education that prepares students for further education and learning for the unplanned. One of the main dilemmas is to do that and at the same time legitimize the goals and the content in pragmatic terms.

These ongoing changes involve new ways of thinking about how to manage education and, also, how goals and aims can be expressed. The conception of *Bildung* that remains from the classical period sees education as built around a specific content expressed in specific texts. The idea of schooling was to reproduce a lost world, not to produce the future. This way of thinking has dominated even modern times; the aim of curriculum design is the articulation of this kind of content in the curriculum and the syllabus and in curriculum materials. Changes in knowledge structures, the rapid growth of knowledge and not least improved access to information and knowledge have meant changes also in curriculum design. In Finland and Sweden new curricula express goals more in terms of concepts, theories and models than in terms of content. As a result, more scope is given to professional responsibility and professional skill, which also means more concrete demands as regards the evaluation of outcomes.

Quality Assurance and Evaluation

These changes in basic ideas about education and the concrete changes in how education is politically governed have given birth to new public discourses on education. Changes in welfare societies, in which the effectiveness of public goods is emphasized, create new demands on education and the outcomes of education. It is in this context that new demands on quality have been made. The quality of education, quality assurance, lifelong learning and the knowledge society have been key terms in the public discourse on education and schooling. Before having a closer look at these concepts, I want to insert a discussion on the concept of evaluation.

The concept of evaluation has to be understood as being part of modernization. It is part of modern society and part of modern thinking.

The society that was formed at the end of the 19th century was a society in rapid transition, tinged with emigration, demographic changes, changes in modes of production, and urbanization. Paid work and the new-born democracies brought about changes in living conditions. A technologically advanced society where rational methods were used seemed to be opened up. The slogan for the World Exhibition in Chicago 1933 condenses this message: "Science Explores; Technology Executes; Mankind Conforms." In this new world it was possible to make choices. But choices require information about alternatives. Products and processes had to be valued.

Educational evaluation evolved from this pattern of thinking, but carried also specific characteristics that developed with the new modern ideas about education. These new ways of thinking about education and its role were considered in the progressive education movement. Dewey's ideas and work are perhaps the best illustration. His pedagogical thinking focused on three concepts: the individual, society, and the practical use of knowledge. The kernel of all pedagogy is the individual. The individual is the hub of the moving wheel. Education must make it possible for the individual to have organized experiences. Modern society was an invisible society, and education was the way to make it visible. Knowledge had to be understood on the basis of its pragmatic values. To understand something, we have to understand how it is used. In this thinking, science was to provide the basis for action. It was important to test experiences in a systematic way. To develop education it was necessary to conduct experiments that could be tested and evaluated. Two years after Dewey was appointed to his chair at the University of Chicago – in 1896 – he started, together with his wife, an experimental school. This experimental school concretized the idea that pedagogical knowledge could be conquered by means of well designed experiments and evaluation.

Evaluation as part of modernization has, as far as education is concerned, two sides. Evaluation was necessary in order to be able to make rational choices. At the same time, evaluation was the way in which knowledge was gained of practice. By means of evaluation it was possible to organize learning possibilities better and better and to choose teaching methods; evaluation was the method to acquire knowledge – the method for creating knowledge out of practice.

These are the basic threads from which the ideas about and aspirations for the evaluation of educational systems were woven.

The 20th century saw a society where education and work were linked to each other, where the labor market and education were intertwined. Salary was related to education. This linkage between education and work made possible the use of education as an instrument for creating a new society and for fostering a democratic citizen. The expansion of education demanded more resources, which in turn increased the demands on evaluation. These demands were mainly formulated after the World Wars. In the Nordic countries, voices for national evaluations of educational systems were raised above all in connection with the school reforms of the fifties, sixties and seventies.

The research on evaluation formed in the seventies was mainly influenced by the USA, but carried also its own specific traditions. The central question in the debate in the Nordic countries was about the pros and cons of the comprehensive school system, and much of the debate concentrated on the question of ability grouping, or to formulate it in another way, on the question of the differentiation between specialization lines. This called for an evaluation strategy in which two alternatives could be compared. At that time, evaluation research delivered answers to how to compare; these were statistical answers to how to compare under non-experimental conditions. This formed a tradition in which comparisons seemed possible irrespective of different circumstances. In Sweden, a major study was carried out with the mission of comparing different types of educational systems – ability grouped or not. The results of these comparisons gave us a basis for the decision to introduce the comprehensive school system in Sweden. When evaluations of this kind were coming to an end, they opened up for international comparisons. While decisions were made on new reforms, the question put to evaluators became more sophisticated. It was no more a question as to which is best, A or B. Rather, the question was: What are the benefits of A or B in relation to the goals.

We can see these changes in educational evaluation also on the international scene. When Senator Robert Kennedy at the beginning of the sixties argued that all federal reforms had to be evaluated, it was in response to public criticism of federal interventions. The consequence was that new fields of research and applied research directed toward explanations of why reforms worked or why they did not work were initiated.

In the beginning, traditional types of evaluation dominated. Evaluations built on the use of tests and advanced statistical analyses. The model had been developed during the Second World War in the Ministry of Defense under the leadership of Robert McNamara. The question approached was: What is the best educational alternative at the lowest cost? We faced the same question in the Nordic countries in the fifties and sixties.

The methods developed in educational research fit well the type of questions addressed when making system reforms. In addition, educational research and educational evaluation were asked for. The construction of the welfare society and the development of social research went hand in hand. Educational reforms and educational evaluation provide perhaps the best illustration.

When the main system reforms were over, the questions addressed changed character. Now the management of education came into focus. How to construct goals and how to evaluate the achievement of goals became essential questions to evaluators. Educational technology gave one answer. An answer which in the Swedish case was illustrated by the MUT project. The model was developed for rather simple educational tasks and aroused strong feelings and criticism. Basic democratic goals lost their meaning when broken down into behaviors. Michael Scriven (1972) even argued with success for goal-free evaluation.

During the seventies, a lot of new evaluation models were born. We had a rather extensive methodological discussion on quantitative and qualitative methods. The case study methodology was developed among others by Robert Stake (1995) and became more than a method: It became a model for evaluation. During the eighties and nineties, a profession of evaluators was established with its own organizations and also with demands for a license.

Bibliography

Andrén, B., Berger, G., & Skjönberg, B. (1998). *Kvalitetssäkring i skolan* [Quality assurance in school]. Stockholm: Skolverket.

Dewey, J. (1899). *The school and society*. Chicago: Chicago University Press.

Dewey, J. (1916). *Democracy and education*. New York: MacMillan.

Dewey, J. (1938). *Experience and education*. New York: MacMillan.

Feinber, W., & Feinberg, E. (1979). *The invisible and lost community of work and education* (Reports on Education and Psychology No. 1). Stockholm: Högskolan for lärarutbildning.

House, E. (1978). Assumptions underlying evaluation models. *Educational Researcher, 7* (3), 4-12.

Karier, C. L., Violas, P., & Spring, J. (1973). *Roots of crisis: American education in the twentieth century*. Chicago: Rand McNally.

Liedman, S.-E. (1997). *I skuggan av framtiden: Modernitetens idéhistoria* [In the shadow of the future: The history of ideas of modernity]. Stockholm: Bonnier Alba.

Oxford Guide to the English Language. (1998). London: Oxford University Press.

Pierce, W. S. (1974). *Pragmatism*. New York: New American Library.

Popham, I. W. J., Eisner, E., Sullivan, H. J., & Tyler, L. (1969). *Instructional objectives*. Chicago: Rand McNally.

Scriven, M. (1972). Pros and cons about goal-free evaluation. *Evaluation Comment, 3*, 1-4.

Stake, R. E. (1995). *The art of case study research*. London: Sage Publications.

Thomas, L. G. (Ed.). (1972). *Philosophical redirection of educational research*. Washington: National Society for the Study of Education.

Tyler, R. W. (Ed.). (1969). *Perspectives of curriculum evaluation*. Chicago: Rand McNally.

Tjeerd Plomp
Faculty of Educational Science
and Technology
University of Twente
The Netherlands

The Potential and Challenges of International Comparative Studies of Educational Achievement

The IEA: What It Is and Does - Its Mission and History

The IEA, the International Association for the Evaluation of Educational Achievement, is the organization that conducts international comparative studies in which educational achievement is assessed in the context of process and input variables. The IEA's mission is to contribute, through its studies, to enhancing the quality of education.

The IEA has developed over a period of 40 years as a co-operative of research institutes, representing at present 55 educational systems (see for example Husén & Postlethwaite, 1996, for a concise description of the history of the IEA). Nowadays, many countries are represented in the IEA's General Assembly by policy makers. The National Research Coordinators and National Research Centres involved in the IEA studies are often among the most prominent researchers and research institutes in these countries; some of them are part of these countries' Ministries of Education, others are linked to universities or are independent research centers. By its nature, the IEA provides a network of institutes and individuals that altogether represent much experience and intellectual capacity. In that way it is a meeting place for policy makers, educators and scientists and researchers.

Over the years, the IEA has conducted many surveys of *basic school subjects*. Most of them have been *curriculum driven*, i.e. a test grid for measuring educational outcomes was developed based on an analysis of the curricula of the participating countries. All these studies included also instruments for measuring school and classroom process variables, as well as teacher and student background variables. Examples are the studies of mathematics and science, reading literacy, civics education, and English and French as foreign languages.

The IEA also conducts studies which are not curriculum based. Examples are the Pre-Primary Project and the Computers in Education Study as well as its successor, the Second Information Technology in Education Study.

At present, the IEA is conducting several studies.

The *Third International Mathematics and Science Study (TIMSS)* has been the largest international comparative study of educational achievement ever made. The TIMSS achievement testing in mathematics and science included:

- 45 countries;
- five grades (the 3rd, 4th, 7th, and 8th grade, and the final year of secondary school);
- more than half a million students;
- testing in more than 30 different languages;
- more than 15,000 participating schools;
- nearly 1,000 open-ended questions, generating millions of student responses;
- performance assessment;
- questionnaires from students, teachers, and school principals containing about 1,500 questions;
- thousands of individuals to administer the tests and process the data.

TIMSS was conducted with attention to quality at every stage of the process. Rigorous procedures were applied to translate the tests, and numerous regional training sessions were held in connection with the data collection and scoring procedures. Quality controllers monitored the testing sessions. The samples of students selected for the testing were scrutinized according to rigorous standards designed to prevent bias and ensure comparability. This monitoring of the quality of the study resulted in marking the countries that did not meet all the quality criteria in the tables with results.

Achievement results of TIMSS have been published by the International Study Center at Boston College (the U.S.); see the references for publications from this study (further reading in the references). Some of these results are summarized and discussed to illustrate the potential richness of international comparative assessment studies.

The IEA was invited to repeat the TIMSS study for grade 8 in 1998 in the southern hemisphere and in 1999 in the northern hemisphere. A number of countries that did not participate in TIMSS were able to join the TIMSS-Repeat study thanks to the World Bank support.

Another study the IEA is carrying out at present is the *Civics Education Study* (CES). This study finished its first phase, the development of country profiles, in 1998 and collected data on schools, teachers and students during the first half of 1999.

Still another study, different in scope, is the *Pre-Primary Project,* a study of child care policies and practices.

The *Second Information Technology in Education Study* (SITES) started in the fall of 1997 with an indicators module (a limited school survey in November 1998). Two other modules have been added, namely a module of international comparative case studies of innovative practices in the use of information and communication technology, and (for the year 2001) a survey of schools, teachers, and students.

The IEA recognizes *two purposes* as far as international comparative achievement studies are concerned:

1. to provide policy makers and educational practitioners with information about the quality of their education in relation to relevant reference groups; and
2. to assist in understanding the reasons for observed differences between educational systems (which serves policy makers' needs, but is clearly among researchers' interests).

In line with these two purposes, the IEA strives in its studies for *two kinds of comparisons.*

The first one consists of direct international comparisons of effects of education in terms of scores (or subscores) in international tests, as is illustrated, as far as TIMSS is concerned, in Table 1 (see below).

The second kind of comparison is concerned with how well a country's intended curriculum ('what should be taught in a particular grade') is implemented in schools and achieved by students. This kind of comparison focuses mainly on national analyses of a country's results in an international comparative context. A typical IEA study deals with grade levels in three populations: elementary education, junior secondary education and senior secondary education.

The IEA was founded as a research co-operative. Initially, it was primarily interested in international comparative studies from a research perspective. In the second half of the 1980s, the IEA started to recognize the increased interest of policy makers in educational indicators. Since then, the IEA has taken it as a challenge to serve, through its studies, also the interests of policy makers. The inclusion of IEA achievement indicators in the OECD's publications is an indication that the IEA has started to become successful in this. The OECD's *Education at a Glance* (1997) presents a number of indicators based on the results of TIMSS. Examples of IEA publications which address relevant policy questions are Postlethwaite and Ross (1994) and Keeves (1996); another relevant source is Kellaghan (1996).

Although it is not necessary for every study to be as exhaustive in size and design as is the TIMSS study, the IEA strongly believes that the conceptualization and design of its studies allows for designing studies which meet the needs of both policy makers and educational practitioners.

Functions of IEA Studies

The relevance of IEA studies extends beyond making just direct comparisons in the form of league tables. The following functions illustrate the importance of international comparative achievement studies (and of educational indicators).

Description: The Mirror Function

This function serves to provide policy makers and the education community with information about the status of 'their' educational system in an international comparative context; this in itself is considered interesting by many. Many policy makers have now recognized that this kind of information is a good starting point for generating questions for in-depth analysis. This can be illustrated with some exemplary results from TIMSS presented in Table 1 (also discussed in Plomp, 1997), which contains achievement test results for science in the 7th and 8th grade.

Table 1
TIMSS – Average Achievement in Science

Eighth Grade*		Seventh Grade*	
Country	**Average Achievement**	**Country**	**Average Achievement**
Singapore	607	Singapore	545
Czech Republic	574	Korea	535
Japan	571	Czech Republic	533
Korea	565	Japan	531
Bulgaria	565	*Bulgaria*	531
Netherlands	560	*Slovenia*	530
Slovenia	560	Belgium (Fl)	529
Austria	558	*Austria*	519
Hungary	554	Hungary	518
England	552	*Netherlands*	517
Belgium (Fl)	550	England	512
Australia	545	Slovak Republic	510
Slovak Republic	544	United States	508
Russian Federation	538	*Australia*	504
Ireland	538	*Germany*	499
Sweden	535	Canada	499
United States	534	Hong Kong	495
Germany	531	Ireland	495
Canada	531	*Thailand*	493
Norway	527	Sweden	488
New Zealand	525	Russian Federation	484
Thailand	525	Switzerland	484
Israel	524	Norway	483
Hong Kong	522	New Zealand	481
Switzerland	522	Spain	477
Scotland	517	Scotland	468
Spain	517	Iceland	462
France	498	*Romania*	452
Greece	497	France	451
Iceland	494	*Greece*	449
Romania	486	Belgium (Fr)	442
Latvia (LSS)	485	*Denmark*	439
Portugal	480	Iran, Islamic Rep.	436
Denmark	478	Latvia (LSS)	435
Lithuania	476	Portugal	428
Belgium (Fr)	471	Cyprus	420
Iran, Islamic Rep.	470	Lithuania	403
Cyprus	463	*Colombia*	387
Kuwait	430	*South Africa*	317
Colombia	411		
South Africa	326		

Source: IEA Third Intentional Mathematics and Science Study (TIMSS), 1994-1995

* Eighth and seventh grades in most countries.
Latvia is annotated LSS for Latvian-Speaking Schools only. Countries shown in italics did not satisfy one or more guidelines for sample participation rates, age/grade specifications, or classroom sampling procedures. The report presents standard errors for all survey estimates.

BEST COPY AVAILABLE

Table 1 illustrates one of the purposes of international comparative achievement studies, namely that of *providing policy makers and educational practitioners with information (indicators) about the quality of their educational system in relation to relevant reference groups of similar nations.* This is the *'mirror' function:* Countries can determine whether or not they like the picture or profile of their country as compared to other countries.

Table 1 just gives "horse race" data, with England in 10th place in the 8th grade (in science) and in 11th place in the 7th grade. Table 2, on the other hand, provides an overview of the countries which are performing significantly better/more poorly than or statistically not differently from England.

Table 2
TIMSS - Mathematics: England vs. Other Countries

Significantly higher achievement:

Singapore
Korea
Japan
Hong Kong
Belgium Fl.
Czech Rep.
Slovak Rep.
Switzerland
Netherlands
Slovenia
Austria
France
Hungary
Russian Fed.
Australia
Ireland
Canada
Belgium Fr.
Sweden

No significant difference:

Thailand
Israel
Germany
New Zealand
Norway
Denmark
USA
Scotland
Latvia (LSS)

Significantly lower achievement:

Spain
Iceland
Greece
Romania
Lithuania
Cyprus
Portugal

This type of information tells policy makers in England and Wales how well their country is doing in comparison with other countries. It also shows that league tables like Table 1 contain limited information and may result in misleading interpretations, as they do not reflect any statistical information.

However, information gained from tables like the ones above does not help policy makers, curriculum developers and educational practitioners to understand why *their educational system is performing as it does, for example, why England (together with Wales) is performing more poorly than many of its EU partners.*

The broad interest world-wide in the TIMSS results illustrates the relevance of this function.

Benchmarking

This function can best be illustrated with an example. Within the TIMSS study, some Asian countries and in Europe Belgium (Flemish) and the Czech Republic have the highest test scores in mathematics. Any country interested in improving its teaching of mathematics can analyze its 'own' case against the Asian countries and/or the European countries with respect to many variables, related to curricular aspects of mathematics and science education (including curricular materials), pedagogical approaches and instructional processes, school variables, teacher background, teacher training (and in-service training), etc. Such analyses may result in proposals for change, although no easy answers can be expected. For such countries, an important question in the next IEA study would be whether their performance will then be closer to that of the reference countries chosen.

'Monitoring' the Quality of Education

One step further than benchmarking goes monitoring: the regular assessing of educational processes on different levels of the educational system with the purpose of bringing about change when and where needed ('informed decision-making'). This function is an example of assessment-led monitoring of the curriculum (but in the case of IEA studies, on the basis of curriculum-based assessment). For this use, *trend data* are needed, i.e. a cycle of regular assessments in the subject areas which are being monitored (such as the IEA and OECD cycle of studies in mathematics, sciences and reading literacy). For this reason the IEA was asked to repeat the TIMSS study for the grade 8 population in 1999.

'Understanding' the Reasons for Observed Differences

Policy makers may want to understand differences between or within educational systems from the perspective of *national* policy-making (this function should be distinguished from the next one: cross-national research).

This function is again one step further than just collecting data for monitoring purposes: It serves ultimately policy makers' needs, but is clearly among researchers' interests as well. To realize this function, information about learning and teaching processes and their inputs as well as in-depth analyses of achievement results in the context of this background data are needed. IEA international comparative studies collect different kinds of background data as well, but the IEA considers this type of analysis an important task of the participating countries themselves as they can best bring up the research and analysis questions which are relevant to their educational systems. Below, an example from Switzerland relative to TIMSS will be presented.

Another good example is the analysis done in the USA of the data of the IEA's Second International Mathematics Study (SIMS) resulting in a monograph, *The Underachieving Curriculum* (McKnight et al., 1989), while an example from TIMSS will be discussed below. Here again no easy answers can be expected as to what measures should be taken to improve education in a country. But this kind of research may lead to policy decisions about changes in education ('informed decision-making'), or to initiatives as was the case in the USA, where the NCTM (National Council for the Teaching of Mathematics) developed the well-known standards for the teaching of mathematics.

'Cross-National Research'

This function refers to exploratory and/or in-depth research on the IEA databases. In TIMSS, this in-depth analysis is still to be done.

Many examples can be found in the IEA volumes. Here I only want to mention two other examples. Postlethwaite and Ross (1994) carried out an exploratory study on the IEA reading literacy database (data collection in 1990-91) in an effort to find indicators discriminating between more effective and less effective schools in reading.

The second example is Keeves' (1996) monograph *The World of School Learning: Selected Key Findings From 35 Years of IEA Research,* in which he discusses, on the basis of all IEA studies conducted until 1994, ten key findings with suggested implications for educational planning.

What Data to Collect: Some Practical and Theoretical Considerations

The question as to what kind of data should be collected in an international comparative assessment study cannot be answered unambiguously. The question is not a trivial one when one realizes that in most IEA studies more than 20 countries and in TIMSS even more than 40 countries are participating. Many participants may differ in the functions they want to concentrate on or the goals they want to reach by means of the study. Some may want to emphasize the description of only a small number of indicators, while others strive for a large number of variables in order to be able to analyze their country's data properly. Besides, according to its mission, the IEA does want to create opportunities for conducting cross-national analyses in order to enhance the understanding of the functioning of educational systems at all levels. On top of this, there is the dilemma between desirability and feasibility: Researchers may desire to collect as much data as possible to be able to do in-depth secondary analyses of a rich database, while the usually restricted possibilities to collect data in schools as well as limited budgets impose severe restrictions on the size of the data collections. So, in this type of studies compromises have to be made between the interests of all participating countries. The IEA is therefore striving for a design and for instruments that are as 'equally unfair' as possible to all participating countries.

For a study to be effective and efficient, a well-thought conceptual framework addressing the issues to be considered in the study is necessary. Almost all of the functions mentioned above need measuring of educational achievement and other educational outcomes on three levels of the educational system:

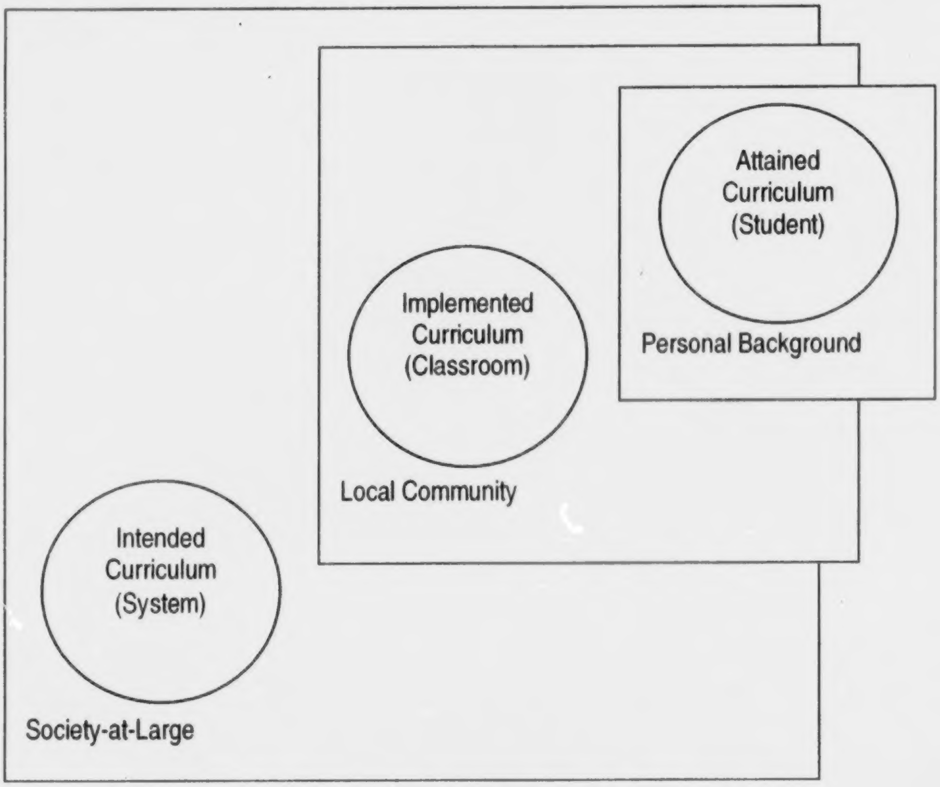
Assessment	**System level**
What students learn	Micro level
What and how schools and teachers teach	Meso level
What the community values (what students should learn)	Macro level

IEA studies address all these three levels by distinguishing between three appearances of the curriculum:

- The *intended* curriculum: What should be taught and learned; can be measured by analyzing documents such as official syllabi, course outlines, text books;
- The *implemented* curriculum: What is actually being taught or taking place in schools and the classroom – the content, time allocations, instructional strategies, etc.; to be measured by means of questionnaires (or observations);
- The *attained* curriculum: What students attain or learn in terms of cognitive skills, attitudes, etc.; to be measured by means of tests.

In, for example, the conceptual model for the TIMSS study, the variables influencing education are seen as "situated in a series of embedded contexts starting from the most global and moving to the most personal one", as is illustrated in Figure 1 (Robitaille, 1993; pp. 26-27).

Figure 1. The conceptual framework for TIMSS (Robitaille, 1993, pp. 26-27).


Attained
Curriculum
(Student)
Personal Background
Implemented
Curriculum
(Classroom)
Local Community
Intended
Curriculum
(System)
Society-at-Large

For more information about the conceptual approach of the IEA, see for example Robitaille and Garden (1992) and Plomp (1992).

In a typical IEA study, many actions have to be taken to collect and provide data and indicators of good quality. These include curriculum analysis; instrument development (including pilot testing, translation, etc.); sampling; the production of instruments; data collection, cleaning and file building; quality control of each component done in the participating countries; data analysis; report writing.

What Data to Collect: Some Examples

In relation to the practical and theoretical considerations discussed above, the questions as to what data should be collected in national and international assessment studies can still be answered in various ways. Again, the answers depend on the functions of the study as well as on the research questions that the study is going to address. On top of that, participating countries may want to use an international comparative study to find answers to some national questions as well. Therefore, the 'what data' question has to be answered for each study separately. Here we will present some examples typical of IEA studies.

Data From What Target Populations?

The choice of the target population(s) is clearly a reflection of the (policy or research) questions in which one is interested. For example, in its cycle of achievement data collections, the OECD collects data from 15/16-year-olds, in order to be able to provide policy makers with a baseline profile of the achievement of students at (or close to) the end of compulsory schooling. On the other hand, in the IEA TIMSS study data were collected (among other things) in the 3rd/4th grade (population 1), the 7th/8th grade (population 2) and the final year of secondary school (population 3), which allows for several comparisons. First of all, the growth between two adjacent grades can be measured. But by including items common to both populations in the tests, also the growth in mathematics and science from grade 4 (elementary school) to grade 8 (junior secondary school) can be measured. In TIMSS, also comparisons between populations 2 and 3 can be made. Moreover, the IEA target populations allow for monitoring the quality of education during compulsory schooling.

Multiple Assessment Measures

In the IEA TIMSS study achievement data were collected in two ways. The achievement tests taken by all students in the study consisted of open-ended questions and multiple choice questions. Besides, a sub sample of the students in populations 1 and 2 did a series of performance assessment tasks in mathematics and science. The performance assessment, which was the same for both populations, was administered in a 'circus' format in which a student completed three to five tasks. The results are reported in Harmon et al. (1997). Table 3 presents some of the results of the performance assessment together with achievement results taken from Table 1 for the countries that participated in both the achievement testing and the performance assessment in grade 8.

Table 3

TIMSS Grade 8: Achievement and Performance Scores for Mathematics and Science

Mathematics				**Science**			
Achievement test (scale pts)		**Performance tasks (av. %)**		**Achievement test (scale pts)**		**Performance tasks (av. %)**	
Singapore	643	Singapore	70	Singapore	607	Singapore	72
Czech Republic	564	Switzerland	66	Czech Republic	574	England	71
Switzerland	545	Australia	66	Netherlands	560	Switzerland	65
Netherlands	541	Romania	66	Slovenia	560	Scotland	64
Slovenia	541	Sweden	65	England	552	Sweden	63
Australia	530	Norway	65	Australia	545	Australia	63
Canada	527	England	64	Sweden	535	Czech Republic	60
Sweden	519	Slovenia	64	USA	534	Canada	59
New Zealand	508	Czech Republic	62	Canada	531	Norway	58
England	506	Canada	62	Norway	527	New Zealand	58
Norway	503	New Zealand	62	New Zealand	525	Netherlands	58
USA	502	Netherlands	62	Switzerland	522	Slovenia	58
Scotland	498	Scotland	61	Scotland	517	Romania	57
Spain	487	Iran	54	Spain	517	USA	55
Romania	482	USA	54	Romania	486	Spain	56
Cyprus	474	Spain	52	Portugal	480	Iran	50
Portugal	454	Portugal	48	Iran	470	Cyprus	49
Iran	428	Cyprus	44	Cyprus	463	Portugal	47
Colombia	385	Colombia	37	Colombia	411	Colombia	42
Intl. Average	**513**	**Intl. Average**	**59**	**Intl. Average**	**516**	**Intl. Average**	**5**

Sources: Beaton, Martin, et al., 1996; Beaton, Mullis, et al., 1996; Harmon et al., 1997

The table illustrates the 'mirror' function of this descriptive data, which may lead to important questions to be addressed by policy makers and educational practitioners in many countries.

BEST COPY AVAILABLE

A number of countries have similar scores for all assessment measures: Singapore consistently at the top and for example Spain, Portugal and Colombia consistently below the international average.

Interesting questions can be asked in, for example, the Netherlands and the Czech Republic. Both countries score high above the international average in the mathematics and science achievement tests, but only close to the international average in the performance tasks. If one values the capability of pupils to do well in performance tasks, then the satisfaction of these two countries with their high scores in the achievements tests should not overshadow the concerns they may have about their average results in the performance tasks.

Some other countries have one result that deviates from a pattern. For example, Switzerland is doing very well in the performance tasks and mathematical achievement, but averagely well in science achievement.

The examples presented illustrate that analyzing descriptive results of multiple assessment measures allows countries to ask questions which may lead to further, in-depth analyses of and/or to discussions about the emphasis and focus in the curriculum.

Background Data

Background data are always collected in IEA studies (see Figure 1). Such data allow us to address research questions as to what factors contribute to good quality education. Another reason for collecting such data is that it allows countries to search for determinants of national results in an international context.

In the IEA Reading Literacy Study, Postlethwaite and Ross (1994) concluded that a large number of background variables had an influence on reading achievement. These were divided into several categories, namely indicators of

- student activities at home;
- school context;
- school characteristics;
- school resources;
- school initiatives;
- school management and development;
- teacher characteristics;
- classroom conditions, teacher activities;
- teaching methods.

Postlethwaite and Ross (1994) analyzed cross-nationally these indicators in the light of the question of what makes a school effective in reading. They found that in order to increase students' reading performance, voluntary out-of-school reading should be fostered among students, particularly during the primary school years; schools should have classroom and/or school libraries; and teachers should emphasize reading for comprehension.

In general, the accumulated experiences gained in IEA studies in combination with the questions to be addressed in a study determine to a large extent what background data should be collected from schools, teachers and pupils.

A Need for National Assessment

International comparative studies can be utilized by a country to study its own educational practices in an international comparative context. In the case of Switzerland, Moser (1997) analyzed, in relation to mathematics and TIMSS, the extent to which instructional practices (child-oriented vs. subject-oriented instruction) and instructional variables (the autonomy of students in child-oriented classes vs. on-task behavior in subject-oriented classes) influenced learning outcomes, not only mathematics achievement but also students' internal activity, self-activity and interest in mathematics. He concluded that instructional practices and instructional variables do not have a significant effect on mathematics achievement, but many effects on other learning outcomes. In the light of the much better results in Japan (a country with a special emphasis on subject-matter instructional practices and on on-task behavior), he concludes that instructional practices in Switzerland can improve in these aspects.

Another example of a national analysis from Switzerland is related to our earlier conclusion that in TIMSS Switzerland did quite well in the performance tasks and in mathematics achievement, but averagely well in science achievement. Ramseier (1997) analyzed possible causes for this and concluded that this can be explained by a discrepancy between the Swiss science curriculum (teaching priorities) and the science section included in the international achievement test.

Most international comparative studies do allow for a limited number of national questions ('national option'). The example of Switzerland illustrates how important it is that countries participating in international comparative studies think beforehand about the national (policy and/or research) questions they want to address by means of such a study; also what typical characteristics of the national system need to be included in the background questionnaires to allow for relevant national analyses.

Concluding Remarks

In the light of the discussion and reflection on the significance of international comparative studies, such as those done by the IEA in order to evaluate and monitor the quality of education, some concluding remarks can be made.

First of all, participating in international comparative studies bears greater relevance to a country if *important reference countries* are participating as well. For that reason, a study like TIMSS has special relevance to the countries belonging to the European Union as well as Northern American and a number of Asian countries.

But from the viewpoint of many countries, important reference countries are not restricted to a geographical region. Therefore, the participation of Argentina and Chile in the TIMSS-Repeat can still be of great importance to these countries, although they are the only countries from the Latin American region.

The IEA type of studies are logistically and methodologically complex studies. An important feature of IEA studies is the *training* of National Research Coordinators (NRCs). This is an essential component of the study, as many NRCs appear not to be familiar with the methodology and especially the specifics of international comparative studies. Another not so tangible benefit of participating in such studies is the development of a *network of researchers and specialists* (in for example, sampling, psychometrics, test development, data analysis, etc.), which can be tapped into when countries develop their own evaluation and national assessment studies.

One important aspect, often overlooked, is the possibility of *linking national assessments* to international assessments. The proper linking of the two will not only increase the benefits a country can gain from investments in assessment studies, but is certainly also cost-efficient.

Another cost aspect is related to the question *what data* should be collected. As we illustrated in the examples above, policy and research questions should be the primary factors determining what data should be collected. On the other hand, when cost factors come in and have too great an influence on what data are (or are not) collected, one runs the risk of limited usability of the data collected. If in the case of TIMSS the IEA had collected only achievement data (which indeed allows for interesting indicators like the ones in Table 1) but no data on schools, teachers and students, a country like Switzerland would never have been able to do national analyses in an international context and would have missed the unique opportunity to address some important national questions. It is often only a small increase in costs which makes the difference between collecting just achievement data or getting a rich data set which allows for in-depth analyses of important issues.

References

Beaton, A. E., Martin, M. O., Mullis, I. V. S., Gonzalez E. J., Smith, T. A. ,& Kelly, D. L. (1996). *Science achievement in the middle school years: IEA's Third International Mathematics and Science Study*. Chestnut Hill, MA: Boston College, Center for the Study of Testing, Evaluation and Educational Policy.

Beaton, A. E., Mullis, I. V. S., Martin, M. O., Gonzalez E. J., Kelly, D. J., & Smith, T. A. (1996). *Mathematics achievement in the middle school years: IEA's Third International Mathematics and Science Study*. Chestnut Hill, MA: Boston College, Center for the Study of Testing, Evaluation and Educational Policy.

Harmon, M., Smith, T. A., Martin, M. O., Kelly, D. L., Beaton, A. E., Mullis, I. V. S., Gonzalez E. J., & Orpwood, G. (1997). *Performance assessment: IEA 's Third International Mathematics and Science Study*. Chestnut Hill, MA: Boston College, Center for the Study of Testing, Evaluation and Educational Policy.

Husén, T., & Postlethwaite, T. N. (1996). *A brief history of the International Association for the Evaluation of Educational Achievement (IEA). Assessment in Education*, 3 (2), 129-141.

Keeves, J. (1996). *The world of school learning: Selected key findings from 35 years of IEA research*. Amsterdam: IEA.

Kellaghan, T. (1996). IEA studies and educational policy. Assessment in Education, 3 (2), 143-160.

McKnight, C. C., Crosswhite, F. J., Dossey, J. A., Kifer, E., Swafford, J. O., Travers, K. J., Cooney, T. J. (1989). *The underachieving curriculum: Assessing U.S. school mathematics from an international perspective*. Champaign, IL: Stipes Publishing.

Moser, U. P. (1997). *Swiss analysis of the TIMSS data*. Paper presented at the annual conference of the American Educational Research Association, Chicago, IL.

OECD. (1997). *Education at a glance: OECD indicators* (5th ed.). Paris: Author.

Plomp, T. (1992). *Conceptualizing a comparative educational research framework*. Prospects, 22(3), 278-288.

Plomp, T. (1997). *International educational research: the case of IEA*. In S. Hegarty (Ed.), *The role of research in mature education systems* (pp. 184-195). Slough, England: National Foundation for Educational Research.

Postlethwaite, T. N., & Ross, K. (1994). *Effective schools in reading: implications for educational planners*. Amsterdam: IEA.

Ramseier, E. (1997). *Task characteristics and task difficulty: Analysis of typical features in the Swiss performance in TIMSS*. Paper presented at the European Conference on Educational Research, Frankfurt, Germany.

Robitaille, D. F. (Ed.). (1993). *Curriculum frameworks for mathematics and science* (TIMSS Monograph No. 1).Vancouver: Pacific Educational Press.

Robitaille, D. F., & Garden, R. (Eds.). (1989). *The IEA study of mathematics II: Contexts and outcomes of school mathematics*. Oxford: Pergamon Press.

Further Reading

Martin, M. O., Mullis, I. V. S., Beaton, A. E., Gonzalez, E. J., Smith, T. A. & Kelly, D. L. (1997). *Science achievement in the primary school years: IEA's Third International Mathematics and Science Study*. Chestnut Hill, MA: Boston College, Center for the Study of Testing, Evaluation and Educational Policy.

Mullis, I. V. S., Martin, M. O., Beaton, A. E., Gonzalez, E. J., Kelly, D. J. & Smith, T. A. (1997). *Mathematics achievement in the primary school years: IEA's Third International Mathematics and Science Study*. Chestnut Hill, MA: Boston College, Center for the Study of Testing, Evaluation and Educational Policy.

Mullis, I. V. S., Martin, M. O., Beaton, A. E., Gonzalez, E. J., Kelly, D. J. & Smith, T. A. (1998). *Mathematics and science achievement in the final year of secondary school: IEA's Third International Mathematics and Science Study*. Chestnut Hill, MA: Boston College, Center for the Study of Testing, Evaluation and Educational Policy.

Eugene Owen
U.S. Department of Education
National Center
for Education Statistics
The United States

Educational Indicators

Introduction

In the 1990s the international assessment community witnessed a dramatic growth in participation in comparative education studies due in part to countries' growing concerns about economic competitiveness, their adoption of standards movements, and a general shift in interest to outputs, not just inputs, of education. This growth was supported by advances in statistical methodologies and the inclusion of policy makers in planning and implementation, and was evidenced in the widespread media attention and uses of the results to review and set education policy. We can only expect that interest in international assessments of education will remain strong and that the education community and the public will continue to pull away, as they began to do in the 1990s, from reporting on the "horse race" to using comparative data to support educational improvement by means of educational indicators.

Indicator Envy

In this presentation I define indicators and try to differentiate them from the research program and express my opinion about how research feeds indicators. However, the indicators program that the OECD has is something quite different. For me, an indicator is information that is often statistical and tells something about a complex system, in this case education.

I tend to say that we in education have indicator envy of economics. I think that part of the social indicators in the sixties and seventies were a big movement in the United States and also in Europe, to try to figure out how we in education could have some of the power and credibility that gross national products, gross domestic products and unemployment rates have in the imagination of policy makers and of the general public. How could we capture some of that for education?

First we tried to take on all of society and create social indicators, and it did not quite work. So we are going at it again in education and trying to be more specific. I think that this is part of the idea of trying to find some few phenomena or factors that all let us know more about an educational system in some particular way. I often like to think in a mechanistic way that these phenomena or factors turned into indicators are something like the dials on your car or in an airplane which let you know how fast you are going and how much oil you are using.

Defining Educational Indicators

We have a general framework to guide thoughts about educational research and this is also the sort of framework that has been guiding the thoughts about indicators. Starting with context, we could say that context here is the general society in which the educational system is operating. Inputs are usually things like financial inputs, the number of students enrolled, and what kind of teaching staff you have. Processes deal with how students are being taught, what the instructional practices are, what other kind of processes are going on in the school, etc. The outputs are often thought to be outcomes, for instance student level outcomes, or how well students do in different subject areas. General education attainment levels in the population, how many are going to work, how many will go on to post-secondary school are also outcome indicators. So this is the framework that is used in indicators work in the United States and the OECD.

The indicator movement of the OECD has been striving for a few key indicators – fifty-four was the idea – to be able to provide basic information to policy makers about the educational system. I think in the life of the INES project over the last years, over a hundred indicators have been proposed at one time or other, not always in one publication, though. One of the problems in this area is how to come up with agreement on what these indicators should be.

The indicators also help to communicate information related to the educational system. I know what I am talking about when I see enrolment rates, when I see expenditures per capita. These tell me something important. They are meant to communicate to the general public. You can publish them in the newspaper, and people will be able to understand them. They are also stable over time. You can track them over time. In other words, their definition does not change from one day to another.

When you have the indicators, the interest goes immediately from broad comparisons to specifically how a system functions. Then finally the public wants to know why, and this leads you into both qualitative and quantitative research. This is why I think that there has to be some sort of complementary system.

I am very concerned about the notion that if we are doing the indicators at the OECD we do not need to do the research. From a U.S. perspective that is not at all the case, since research is the foundation we really need, while indicators are in some ways a communication tool, they are not the end you are seeking. The understanding is what you want from the research, but you also need to be able to communicate certain aspects of that research quickly. When people say "Well, we are spending now all the money on indicators and we don't need to do the research", this to me is a very disturbing trend. We are working very hard on keeping the TIMSS notion alive, the IEA alive and making sure that we continue having rich research with all of the variables that we are interested in. Somebody has to do the research from which we draw our indicators. Otherwise you have a very, very sterile system that really will stop telling you what you want to know. Yes, I am very concerned about the movements to reduce research efforts.

Developing an International Indicator System

There are a multitude of issues involved in doing indicators or developing an international indicator system. As you can imagine, in an international context, what one country wants to know about is often not what the next country wants to know about. I feel one of my most interesting discussions was with one representative of the Nordic countries. In the U.S. the notion of *excellence* is a very important topic. The Danes, however, do not want to talk about excellence. From our conversations it seems that excellence is a very divisive topic. The Danes want to talk about doing well for everyone. When you start talking about excellence, they think you are cutting out certain people, so they do not like to discuss that topic at all. To a North American it is shocking that people do not want to talk about excellence.

Thinking about coming to agreement about what indicators we want as an international group is no accident. I think that the U.S. is the chair of the student outcomes group in that work because the whole notion of the "horse race" talks about how well we do in mathematics, how well we do in science, how well we do in reading. This is a very American inclination. But when one of our national goals is to be first in the world in mathematics and science and as the twelfth grade results show we have quite a way to go, this creates a dilemma.

The U.S. is really particularly interested in these student outcome indicators, and many of the people I am working with are also interested in other kinds of student outcomes. We are calling these "cross-curricular competencies" – those skills and abilities that are not necessarily subject-bound, but help to lead to success in later life. These might not be measured as mathematics, science or reading or something as straightforward as that.

In addition, there is the area of what contextual variables are important. I think that it is particularly difficult to determine what cultural variables are important and should be measured to really understand why the system functions the way it does.

Measurement is another problem that we have. When we talk about outcomes we are talking about test scores. Do we want to look at these cross-curricular competencies in the same way or do we want to have another way of looking at them? People say we need to look at some domains within mathematics and so we get down to looking at how people do in calculus, geometry, numbers, and operations. My question is: Does that give me enough information or do we want to have something else? People then want to look at the individual item to understand student responses. So it is a question of both the *complexity* and *accuracy* of your indicators: What is simple enough, what is clear enough, and how well do you *measure* what you are looking at? Are we really measuring what the world thinks is mathematics and science, or is this a particular idea which comes from one particular cultural group or another?

As for the Adult Literacy Survey for instance, there is a question one might ask: Is it a particularly Anglo-Saxon way of looking at what adult literacy is? The French have raised that issue with us: 'Well, we do not do this – we do not look at adult literacy in the same way in France as you might do in Canada and in the U.S. Even though some of the Canadians speak French, that does not matter, they are not French and they do things the way the Americans do.' So it is the question of how we measure different complex contexts. Do we want to include, for example, across countries private and public schools? In some countries private schools are not important, in others they are, and in some countries the government pays for private schools.

A big question we have right now is about teacher qualifications and how teachers are trained, what they get to teach, particularly in the case of U.S. mathematics and science for the twelfth grade. We found out that about fifty-five percent of our students taking physics were taught by teachers that did not major in any science. We actually had people learning physics from people that had never really studied physics in any significant way.

Multiple Audiences

Another issue is the level of the educational system for which we want to have an indicator system. We are talking here about an international indicator set, which is quite different from a national indicator set, which I would say is quite different from a local indicator set. At the other extreme, in some ways you can think of a report card to home to parents as a personal indicator set of how Charley is doing and of the personality that Charley has. So it is necessary to determine at which level you are pitching the indicators. Often at the OECD level there are a lot of arguments about indicators that people want that are particularly important to their country but might better be served in a national indicator set than in an international indicator set.

Multiple audiences is another issue that needs to be addressed. We want something for the general public but is that detailed enough for people that have to make decisions about the educational system? So, again you have the difficulty of different levels of indicators and who the indicators are intended for.

I think one of the thorniest issues is: How many indicators do you really need? If you drive a simple car you might have a gas gauge and a speedometer. If you drive a slightly larger, more advanced car you will have a number of dials to accommodate to tell you how many revolutions per minute. You might have a more dynamic oil-gauge. You might need pressure, you might need something that tells you the temperature of your car and so on. How many of those do you need for something as complex as an educational system? Do you need as many as in an airplane? If you think of the educational system as difficult to drive, as difficult to manage or steer. The French have the notion of steering or piloting *(pilotage)* of an educational system. For such a view of *pilotage*, how many indicators do you need? When do you know you have enough information to do a good job in understanding the educational system particularly if you want to use the indicators to make decisions.

Progress in Educational Indicators

What kind of progress have we made in indicators in the last 35-40 years? There have been a spate of national indicator reports. It is fascinating to look at them to see what different countries choose as indicators. I have a couple of examples of indicator sets: one from the U.S., which is a very decentralized system with lots of local control and state control, and the other from the French system, which is very centralized.

In the U.S. the basic responsibility for education is not at the national level; it is at the state level. It is really at the 50 state levels that they really need to pay attention to indicators and to know what is going on. What we try to do is provide comparisons. If we are going to make international comparisons, we often do them with the data for the individual states as well as for nations, and we look at some general background, things that we would call context variables.

We also have indicators of student achievement and attainment. We look at school completion rates that we have at different levels and the mathematics achievement. We actually did an experimental linking study: We projected an earlier mathematics study, using the national data that states had on their mathematics performance, asking how they would have done if they had taken the international test. So this was a way of giving state-specific scores for example on something like TIMSS. Then we tried to figure out how they would have done compared to other countries. We had states that would have done as well as one of the high-performers, almost as well as some of the high-performing countries such as Singapore, and some that are below the poor-performing countries as well – but it shows the variation among the U.S. states compared to other countries. The labor market outcomes, for their part, are looking at unemployment, education earnings and so on. This is the regular indicator factory. You can generate more finance indicators quickly because of the practice in the field of economics to do indicators. This is also something that policy makers are very, very keen on having, so as to know how much they are spending and then looking at it in conjunction with the other indicators.

The primary publications by the French are a couple of indicator reports: One is called *L'état de l'École* and the other one is *Géographie de l'École* with *L'état de l'École*, however, being the one that they have as their flagship publication. It has a similar range of indicators for the indicator system as INES.

For the OECD, *Education at a Glance* is the primary OECD indicator report. Let me give you an idea of the kind of indicators that it contains. It begins with context, such as the demographic and social characteristics of the countries that participate in the project. Then inputs: the financial and human resources that are invested in education, looks at both the human capital and the financial indicators. We also have access to education, participation, looking at who goes to school, where and how long they go for tertiary education, and then a little bit on the organization of the schools. This system contains process variables, such as class size, reports on the staff ratios and how the students report the time, how they are using the time they are in school. Finally, student achievement indicators having to do with Network A. These are all generated from the TIMSS study. So that is the progress.

Culture-Sensitiveness

Most indicators are culture-sensitive; even those that you think are simple are really culture-sensitive. For example, school expenditures are really very difficult when you count your national expenditures on education, because what you get counted as education in one country is paid for by the Ministry of Health in another. So the U.S. looks as if it had very big education expenditures, but much of these expenditures are paid for by other entities in other countries. So the student has the benefit of all of that expenditure, but it is not counted as *education* expenditure. For example, transportation is counted in the U.S., because we buy school buses and send people, but in lots of countries it is paid for by the Ministry of Transport. That is a simple example.

Getting to the large picture of the cross-curricular competency issue, the whole problem of what it is that we want our children to be able to do when they get out of school, is a very thorny issue. You start getting at very deeply held beliefs about the purposes of education when you start talking about outcomes and measuring outcomes that are not as simple as mathematics, science, reading. For example, when you start talking about what *solving problems* is about you run into difficulty. We had a pilot in problem-solving from a North American perspective and it literally made my Swiss counterpart angry: How could you be so silly to say this was problem-solving? To him it was too much. He did not want to do it and he did not do it in Switzerland because the whole notion of problem-solving from an American point of view was too embarrassing. Also, I would say the notion of *self-concept* in the U.S. is something that is really problematic. There was a whole range of self-concept items that were perfectly acceptable in a European context, but I would have been fired the next day if I had had them administered in the U.S.! So there is a whole range of things that are very sensitive when it comes to student outcomes.

The whole notion of a school is problematic when you get to the upper secondary level. In Germany, are children in the dual system going to school? They are getting *educated,* but what about their participation? There are some *Kantons* in Switzerland where they literally do not have principals. Teachers organize things, but then even school organization – when you do indicators of how schools are organized – is also culture-sensitive! When you go further afield, outside the OECD countries, some of the arrangements are very, very different. In Brazil, for example, they have state schools and municipal schools. The state schools have to follow rules that are set down by the government, and municipal schools can just pop up and teach whatever they want to. So how do you make sense out of this kind of system for an indicator report?

Importance Across Cultures

Coming to agreement about what you really want to know is a difficult task within a country or across countries. What is the set of indicators you want? What do those dials look like? What is it we need to know about education? What is important to communicate. I think that in a national report that is to some extent idiosyncratic to each country, but internationally we need to come to some set of what is important across cultures.

First of all, context. What are the things you need to know? Equity is one of them, as an example. But how do you create a social economic status variable across cultures, across countries and how do you measure it – it is very difficult to get this indicator in a survey cross-nationally. How do you set the context – even nationally sometimes?

Secondly, outcomes are very important. They remind me of de Saint-Exupéry who wrote in *The Little Prince* that the most important things are the things that you cannot see. Sometimes I feel that way about indicators. I can read a book of indicators which is very informative and very helpful and gives me a good picture, but I think what I really want to know, the dial I would really like to have I cannot see. And this, I think, is particularly so in outcomes. For some of the outcomes we just do not have a measurement system yet. Still, you have to be able to measure them, in order to have an indicator, and I do not think we can develop all the measures. This is a continuing challenge, a continuing problem. We are trying to work on the CCCs, cross-curricular competencies – such as problem-solving and working in groups, using technology – competencies you do not necessarily learn in any one course, but as we know, information for the idea of production rather than reproduction of knowledge. Those are the kind of outcome indicators that we are interested in, as opposed to the reproduction notion.

Thirdly, teaching. What are good indicators of teaching? We did a video study in the U.S. to look at some of this with TIMSS teachers, and we also looked at their survey answers. What teachers say they are doing and what people looking at them say they are doing are often quite different things. So we have the whole problem of self-report for process, for really getting at teaching. In many ways when you start talking about education, this is the heart of the matter. What teaching is like is really the black box in all of the indicator programs.

And finally, content. I think that the IEA should be congratulated for keeping this up and wanting to look at the content and how it is taught – the sequencing of what children are taught, what they are presented with. I would say the whole opportunity-to-learn issue and indicators of that kind are still in their infancy.

I just want to show you a couple of examples showing what the U.S. did along with TIMSS to answer some of these questions about teaching that came out of the observation and also the content.

We had mathematics experts across national groups from at least 4 – 5 countries looking at the content of the mathematics classes that were videotaped in three countries – in Germany, Japan and the U.S. – looking at what they would say was the quality of mathematics content in the classes that they saw. What would they say, were they high-, medium- or low-level content, with a consensus approach? This was quite an eye opener to the U.S. and Americans: In the sample of teachers, they could not see any content that they would say was of high quality in the classes. It is not to say it does not exist, but in the U.S. it is a rare event obviously; it did not show up in the sample of teachers for TIMSS.

There was another evaluation done by looking at videos. It was about whether teachers just stated concepts, or whether they developed concepts. When teachers were interviewed in Germany, the U.S. and Japan they were asked what the purpose of their lessons was. In the U.S. and to a large part in Germany the purpose was to get the students to solve problems. In Japan it was to understand the concepts. They did not worry themselves about solving problems; their goal was to get the students to understand the concepts of mathematics. There were a high number of classes where mathematics concepts were developed – almost 80 percent of the time in Japan and Germany and very little in the U.S. About eighty percent of the time they just tell them, they do not develop them. What does this mean for the way people are coming to mathematics in the 8th grade in the U.S? Indicators of this kind drawn from a rich observational research study, providing the third level of indicators, were very good at indicating to the American public the real differences in what goes on in schools in the three different countries.

The other thing that indicators of this kind are very good at is that if you know particular myths in a country that have been accepted as educational truths, you can present data to counterbalance them. In the U.S. we have always been told that our school year is shorter, that children in other countries are actually getting more mathematics and science, and what we just need to do is to keep our children in school longer or have longer school years. When we actually looked at the number of hours the children were being exposed to mathematics and science, they were not less in the U.S. than in other countries, because of the way the curriculum was arranged, the way the school year or the school day was run. Doing *more* of something you are not doing well is not particularly helpful, and so this was another way of killing a myth in the U.S. Somehow children need to have more mathematics and science instruction in order to do well, but that was not the difference between what was going on in Germany, Japan and the U.S.

So this was again looking a little deeper, having something to say about the process of education and being specific about it. This can do a lot in changing the nature of debate, and that is what I see the indicators are doing: helping in forming the debate of where to look, of where to go into more depth.

THEME 2

MERGING NATIONAL AND INTERNATIONAL CONCERNS

Erkki Kangasniemi
Institute for Educational Research
University of Jyväskylä
Finland

National Viewpoints on International Evaluation and Research

To provide some background to the viewpoints presented below, I might mention having been involved in the Second International Mathematics Study of the IEA (IEA/SIMS) and in the OECD educational indicators project INES, Network C, as a national researcher. Some comments will be made on international research and evaluation work based on my experiences in these studies, and particularly from a national point of view. This commentary will be neither exhaustive nor systematic, but will bring up some aspects that relate, for instance, to the nature of international co-operation and its effects on our national research and evaluation work.

First of all, it can be claimed that international projects, such as the IEA achievement studies, represent, by their basic approach, the scientific concept of evaluation. They have a theoretical model which is based on a multilevel system of education and the curriculum (Travers & Westbury, 1989), recognizing the role of evaluation in schooling. Such modeling also provides research work with its practical framework.

On the other hand, being based on administrative and policy interests in educational evaluation and in decision-making concerning the development of schools, the OECD project on educational indicators corresponds more closely with an administrative conception of evaluation. It is therefore no wonder that the national school administration in Finland has shown greater interest in the indicator work than in the IEA studies, at least if we look at its interest in organizing and funding our national activities. In particular, the Finnish Ministry of Education has set up and funded a national working party to deal with the OECD indicator work.

The IEA, unlike the OECD, is not an organization for formal intergovernmental co-operation, but has rather evolved around issues of comparability, description and explanation. The IEA studies have therefore been more in the interests of our research institute (the Institute for Educational Research of the University of Jyväskylä), which has acted as the National Research Centre in the IEA studies. The Institute has accepted national responsibility for conducting the studies, and the Ministry of Education has granted financial support for this purpose. It must be recognized that the IEA studies have provided models for the work, whose results have raised wide interest in assessing the Finnish educational system.

IEA Studies and Finland

Up to 1998, Finland has taken part in altogether 12 international evaluation studies organized by the IEA in which our Institute has served as the National Research Centre, and one (the Pre-Primary) study where this duty was delegated to another institute. These activities started back in the early 1960s, in terms of the first Pilot Study, and continued with the First International Mathematics Study (IEA/FIMS), followed by the major early effort: the Six-Subjects Study at the turn of the 1970s.

Participation in these (two) international evaluation studies has provided the Institute for Educational Research with good and useful experience and expertise. It gave us potential to conduct national level assessments since the 1970s, in efforts to evaluate the progress of our concurrent school reforms, and to develop school and its curricula further. Thus, along with international contacts, the Finnish researchers gained experience and expertise which created conditions favorable to national research and evaluation work in general.

National Evaluation Agencies and Institutes in Finland

In Finland, educational research and evaluation have been done at universities and at their research institutes. As far as national level evaluation studies are concerned, the Institute for Educational Research has had a central role in Finland. In addition, various university faculties and departments have hosted research projects dealing with education, such as, for example, those commissioned by the National Board of Schools in the 1970s and 1980s. At that time in Finland, research on schools was not done by the Ministry of Education nor by the former National Board of Schools – presently the National Board of Education – but by the academic society, i.e. universities. Besides expertise and capacity issues, this strategy was chosen because of possible reliability and credibility concerns. Results of school research conducted by the central government were thought to be regarded as less reliable than those coming from the academic circles, which are considered perhaps more objective and neutral, without vested interests. This aspect has been brought up, for instance, in a doctoral dissertation (Männistö, 1997) and related discussion on these issues.

The Reliability Issue

In each of the IEA studies the expertise of the International Specialist Committee has been outstanding, and even decisive in facilitating high quality research and reliable results. The international nature of IEA studies, and also the fact that the IEA is not about intergovernmental affairs, have protected the research from national-political interests and from compromises for political expedience.

In this context, the issue of reliability reminds me of an incident at a meeting in Toronto, during the Second International Mathematics Study (IEA/SIMS). The question of the publicity of the results suddenly turned up. The question was raised by some members in the assembly who were representing the Ministries of Education in their respective countries. For researchers it was self-evident that the results were to be published. But for some of the ministerial representatives this issue was threatening, so they proposed that if their results turned out to be poor, they should not be published. As an alternative they suggested that in any publications, an obscure code system should be used instead of naming educational systems or countries, so that it would be impossible to put one's finger on any particular system/country. The researchers were amazed at this idea of not being open with the results. As we now know, the results of the study have been published in three international reports and in several national reports. In addition, they were addressed widely by the press in several countries. – It is desirable that when launching an international study, everyone is aware of and accepts the basic rules of scientific research and heeds sound research ethics, which includes publishing the results, as well.

National Benefits of International Co-operation

When joining an international study, the question of its *national benefits* is always brought up. As benefits we could name, for example, the acquisition of new knowledge, research experience, and international professional contacts. Countries that have participated in the IEA studies are typically interested in follow-up type of research on their school systems. Here, the target population becomes a key issue. Is the target population *relevant* from the viewpoint of the national school system? Secondly, is the target population *the same* as in the earlier study, offering opportunities to contrast the results with the earlier ones, and to evaluate the educational outcomes of the national education system at its different stages of development?

For Finland it has been important to participate both in the First (FIMS) and in the Second Mathematics Studies (SIMS), and similarly, in the First and Second Science Studies (FISS and SISS). The target populations were the same in both of the mathematics studies, and also in both of the science studies (although not across these studies). Furthermore, the studies contained some tasks in common – so-called anchor items. This meant that the studies were partly repeated in the same target populations, but at different developmental stages of the Finnish educational system. In this way we got an opportunity to evaluate and compare the teaching and learning achievements for the same populations in the 1960s and in the 1980s. Back in the 1960s, Finland had what was called a parallel (or binary) school system, but in the 1980s, as a result of a school reform, we changed over to an integrated comprehensive school system, uniform for the entire age group.

When the comprehensive school system was being introduced, people were worried about the standards of education. The above-mentioned IEA studies, among other things, have made it possible to compare teaching and learning achievements in the older, parallel system and in the newer, comprehensive school system. By means of these studies and comparisons, researchers have been able to inform the pupils of the 1950s and 1960s, i.e. today's decision-makers, who would otherwise cherish their golden memories of and beliefs in the good old days when school was better and learning achievements just excellent. We can say that the first two IEA studies in mathematics and in science were unique in the Finnish situation. They are still the only studies yielding comparable research evidence of the Finnish school system at its different stages. These IEA studies are therefore valuable for us also in the historical sense. The timing of these studies has matched well with certain developmental stages of our school system.

The unfortunate fact that we were not able to participate in the Third International Mathematics and Science Study (TIMSS) by the IEA has been remedied, however, as Finland has been involved in the current repetition phase of the study (TIMSS-R). The data collection took place in 1999, and once again we have research data and results available concerning the teaching and learning of these subjects in a new curricular situation.

Curriculum Analysis

National school systems tend to have certain characteristics of their own. International studies seek to take these national characteristics into account to the extent practically possible, and to ensure that the research and its results are as valid as possible for each school system. One such component in the IEA studies is curriculum analysis. It seeks to define certain essential input factors of schooling. By means of national analyses of objectives and content, the overall frame for international item pools can be determined, from which the items for common instruments can then be sampled. This will increase the content validity of the instruments for each country and school system.

Again, the curricular situation in Finland has changed. Back in the 1970s and early 1980s, we used to have a centralized national curriculum which determined the inputs, objectives and content of teaching in a uniform manner for all comprehensive and senior secondary schools. At the time it was relatively easy to do an analysis of the national curriculum: There was only one, detailed and comprehensive, national curriculum for each target population to be analyzed. The present situation is different: We have a national *framework curriculum*, written on a rather general level, on the basis of which schools prepare their own, more detailed curricula, often with particular emphases. As may be expected, the school-based curricula vary in format and quality. In some schools they are written plans, in some others they may be partly written, partly taped, etc.

How should one go about analyzing the national curriculum in this situation? Should we analyze the framework curriculum and all the school-based curricula of the schools in the sample? Or should we analyze the framework curriculum and the textbooks available? For each subject there are usually three or four textbook series as alternatives from which the schools can choose. As we all know, a textbook tends to be an operationalization of the curriculum, telling what teaching should be like content-wise. Are textbooks becoming a significant part of the curriculum in this new situation? Are the textbooks complementary to the general-level framework curriculum? Curriculum analysis may in this situation end up being very textbook-oriented, which of course would not be curriculum-independent analysis, either.

We have to bear in mind that the IEA studies include also opportunity-to-learn (OTL) assessments, done by each teacher of the sampled classes, to complement the overall curriculum analyses. Thus, the OTL assessments reflect the implemented curriculum and relate to the validity of the study from the national point of view.

The Role of National Expertise

The curriculum analyses in the international studies are prepared by national experts in accordance with an international framework. The contribution of the national researcher and the advisory group should ideally be used in the international studies also when it comes to classifications or interpretations concerning the national school system. However, the editor of an international report may sometimes, relying on his or her own judgement alone, make questionable or even misleading interpretations of a national school system. For instance, in the Second International Science Study, the editor of an international report, overruling the national expert opinion, interpreted a group of students as being specialized in a particular subject. Yet, these students were merely studying this subject as one subject among many other subjects, all of them as a uniform course, and they were studying it only a couple of hours per week. The interpretation was made, however, even though the national researcher could not regard them as specialized in the subject. In cases like this we should accept that the national researcher makes his or her judgements within the international framework and on the basis of his or her acquaintance with the national curriculum. Of course, the national researcher has to justify and provide valid arguments for such interpretations so that the author understands them and is convinced that the argumentation is well-grounded.

There are other cases, too, when national researchers have found it difficult to have erroneous information corrected at the proof-reading stage of the report manuscripts. Such misinterpretation seems particularly likely when it comes to curriculum analyses. It is unfortunate and unacceptable if incorrect judgements remain and get published in an international report despite the national researcher's efforts to correct them.

Educational Indicators

As already mentioned in the beginning, educational indicators are meant to serve policy-making. Educational policy makers want information about the functioning of the educational system, for instance, for the purposes of planning and resource allocations. While educational policy makers content themselves with system level indicators, the authorities in school administration call also for school-based indicators.

The educational indicators of the OECD yield information at the system level about the economy, processes and achievements of education as well as about attitudes toward education. For each of these components, the aim has been to produce a number of indicators reflecting essential aspects pertinent to schooling and education. The significance of the indicators for educational policy, the meaningfulness and attraction of the pieces of information conveyed, however, depend ultimately on the needs and views of the user.

The OECD indicators are system level indicators. There are needs for indicators at other levels as well. Today, in our national education policy the emphasis is on market-orientation and neo-liberalistic principles. In this regard, national school authorities are sometimes calling for school-based indicators which would enable comparisons between individual schools and make them compete with each other, while introducing what is called steering by information. By the same token, in the market situation, parents may also wish to get information about the quality of the schools in the area. This direction may lead us into national and regional work on school-based indicators in Finland. Already, such indicators are applied to universities, for the above-mentioned purposes. For other schools too, indicators could possibly be used as criteria for quality and for resource allocation. However, it will not be easy to reach an agreement upon the principles to be used in the allocation. Should the resources be used as *rewards* where the indicators show success already, or should we direct more resources to those schools where the indicators look poor, so that they could *improve* their results with the extra funding? Exclusively quantitative indicators may prove very harmful to the qualitative development of schools in the context of competition. In such circumstances the schools are tempted to strive for the quantitative goals in every possible way. That kind of competition would not be likely to motivate and encourage qualitative development.

The recent delegation of power to schools and municipalities has meant a clear change in the Finnish administration culture. Now that the central government no longer provides schools with a curriculum or administrative directives, the central government is looking at the situation from the accountability point of view and taking interest in indicator-type information on how Finnish schools are functioning, e.g.: How have the municipalities, as owners and maintainers of schools, directed their state subsidies to school expenditures, when compared to the earlier practice? Has the course of action been in line with their adopted objectives? and so forth. As the planning and decision-making is no longer centrally directed, the central government is increasingly interested in the processes going on and results gained in schools in the light of the various indicators. This is related, among other things, to the current discussion in Finland about possible school-leaving examinations for comprehensive school.

As for international comparisons, although the inevitable conceptual and practical compromises impose certain restrictions, the OECD indicators still yield to a great extent comparable evidence. In general, the indicators have shown considerable variation from one country to another. The process indicators, for example, reveal that for the lower stage of the Finnish comprehensive school (or primary school, if you wish) there are characteristics that may be highly contextual, bound with time and the situation. These include factors such as internal interaction in the school, and discussion between the school and homes about values and objectives, which phenomena are directly affected by the concurrent curriculum work in schools (Kangasniemi, 1997; OECD, 1997). The process indicators also show how we go about with evaluation matters in Finland. Evaluation is still seen in primary schools as student assessment, rather than as a means to evaluate and develop teaching and promote learning. The indicators also tell us that there are few connections between the lower and upper stages of comprehensive schools – they have separate locations and functions etc.

The achievement indicators produced by the OECD have so far been mostly based on the research data collected by the IEA. These indicators have sought to rank the school systems of different countries on the basis of their learning results (OECD, 1996, 1997), which is typical of the indicator thinking. Having been actively involved in several IEA studies and thus acquired similar information earlier, these indicators have not given us much new food for thought so far. While the IEA studies have, in addition, provided information supplementary to achievement indicators, they have offered a broader base for studying and interpreting these indicators. It seems that the contextual information offered by *Education at a Glance* cannot be related to student outcomes with equal confidence.

As Finland was not a participant in the IEA's Third International Mathematics and Science Study (TIMSS), the OECD achievement indicators in 1997 were left blank for Finland. This was noticed here, and the Finnish Ministry of Education, for one, has regretted our absence from TIMSS. This can be taken as an indication that with relation to the indicators, appreciation of the IEA studies has been considerably enhanced among educational policy makers and administrators in Finland.

Feedback and Utilization

It is said that knowledge is power. This is partly true. Because the knowledge must also be used, the feedback obtained put into good use. Only through its utilization or wielding may knowledge become power. In my opinion, we should increase national discussion in Finland, in order to take full advantage of research knowledge and increase its appreciation. It is not enough to publish the study results; the results need to be refined and turned into practical measures. Furthermore, when informing people about the results of studies and evaluation projects, the researchers ought to see to it that the material published includes also information on certain context, input, and process factors involved, and not only the outcomes as such. This is to counteract the prevailing principles of accountability which lay perhaps too much stress on mere school achievements, and therefore they are often published at the cost of information about important frame factors. This is not to say that radical actions are needed.

On the evidence obtained from international studies, I think we should examine, each within our own educational system, the similarities and differences between the functioning of the system and the objectives set to it. In so doing, it should be accepted that each national education system can have a profile and special characteristics of its own, and universal conformity is not the goal.

References

Kangasniemi, E. (1997). OECD:n prosessi-indikaattorit ja ala-asteen koulujen toiminta [OECD process indicators and practices in the primary schools]. In R. Laukkanen (Ed.) *OECD-maiden koulutuspolitiikan analyysi* [An analysis of educational policies in OECD countries] (pp. 28-42). Helsinki: National Board of Education & Paris: OECD.

Männistö, Y. (1997). *Kouluhallitus koulututkimuksen rahoittajana ja tekijänä* [The National Board of General Education as the financer of research and the utilizer of research findings] (Tutkimuksia). Helsinki: University of Helsinki, Department of Teacher Education.

OECD. (1996). *Education at a glance: OECD indicators.* Paris: Author.

OECD. (1997). *Education at a glance: OECD indicators.* Paris: Author.

Travers, K. T., & Westbury, I. (1989). *The IEA study of mathematics I: Analysis of mathematics curricula.* Oxford: Pergamon Press.

Pirjo Linnakylä
Institute for Educational Research
University of Jyväskylä
Finland

The National Intertwined With the International

"Evaluation is an investment in people and in progress"
(Guba & Lincoln, 1989, p. 2)

Growing Interest in International Evaluations

A small nation, whose most important resource is the people, especially the children and the young, cannot afford negligence when monitoring the development of education. It is not enough to just think we are doing fine – we must also do so – and therefore we must first know what we are good at, nationally and internationally, and what needs to be improved (Clarkson, 1994).

Our children and youth are already living in a world where students and the workforce are moving across borders not only as interrailers or tourists but as new active citizens of the world. The barriers to international interaction have been considerably lowered by the rapid advance of and easier access to information technology. This trend is but speeding up. New technologies have also transformed the educational environment, making international interaction and co-operation an everyday part of students' life and learning.

The people of today, especially the young, are truly interested to know what kind of education they will have, on which to lean when leaving for the world. And in general, what kind of education will be required in our society which is becoming more international and in the outside world? These are questions which cannot be ignored by educational policy makers or researchers. We need evaluation information on many issues: What is the level of our national competence? What kind of desire for learning and confidence in our competence do we have? What about our communicative and interactive skills, both face-to-face and through networks? How do we face other cultures and multiculturalism? Does our educational system prepare us for lifelong learning and for future employment? Is it capable of fostering people's personal growth and active citizenship as individuals and of providing sustainable and balanced development to the nation as a whole?

The Multicultural Concept of Evaluation

In the context of international evaluation projects in recent years, a point of shared interest has been the view of education as a national investment and a source of added value. There is now a global interest in the structure, value basis, and functioning of educational systems, as well as in the standard and quality of educational processes and their outcomes, i.e. the competencies produced. The interest is similar in the equivalence of qualifications and in the social consequences of education, especially with regard to employment. Along with the development of technology, the discrepancy between educational outcomes and the competencies required in working life has increased the need for international co-operation, and in particular, for national and international evaluations of competence.

Although evaluation has traditionally not been considered real scientific work, it is gaining an ever stronger foothold as a scientific discipline of its own (Scriven, 1991, 1994). It is regarded as an applied transdicipline, the strength of which derives from its societal relevance – both nationally and, increasingly, internationally. Today, the theoretical and methodological demands made on an evaluation scheme are high. Evaluation activities resemble research work. Evaluation must have a sound conceptual basis and a theoretical frame. Furthermore, the data for evaluation must be collected systematically, efficiently and economically employing different research paradigms and methods and so that the information obtained is reliable, credible, generalizable, as well as up-to-date, content-wise and context-based. (House, 1995; Norris, 1995.) Consequently, the conclusions drawn on the basis of the data should extend our knowledge and lead to relevant improvements. A special emphasis has to be put on the social and individual effects and consequences the evaluation may have (Messick, 1992).

The prevailing conception of evaluation – the fourth generation of evaluation (Guba & Lincoln, 1989) – accordingly emphasizes a constructivistic and socio-cultural approach, where different values and principles have to be negotiated with a view to forming a shared framework for evaluation, and moreover, preferably so that the values and informational needs of different interest groups – both individuals and nations – are taken into account. A special emphasis is laid on the rights and protection of the target of evaluation (Norris, 1995). Cultural pluralism is also attached to an eclectic and interdisciplinary approach. Evaluators must have content area knowledge accompanied with expertise in evaluation theories, methods and practices as well as psychological and societal understanding. Multiculturalism encompasses multiple methodologies, both quantitative and qualitative approaches, and integration of different theoretical frameworks and evaluation procedures (Guba & Lincoln, 1989; House, 1990; Scriven, 1994).

In international evaluation schemes, this means matching up quite varied national values, visions and standards in an open-minded fashion. It is in this way only that evaluation can be ethically justified and operationally sensible, promoting both national and global interests. As any other type of research, evaluation is valid only if its results can help learn more about the target, enabling conclusions with relevance to educational objectives and underlying values, and leading to viable solutions for development (cf. *Standards for Educational and Psychological Testing*, 1985, p. 9).

Making the right decision on developmental actions, in turn, calls for expertise which is rooted in a national context, in national realities and visions. After all, the validity of evaluation also lies in its consequences – in educational development work and in educational research.

The Added Value From International Evaluation Schemes

International evaluations can reveal, more clearly than national ones, the special characteristics of a particular education culture with respect to its context, processes, learning achievements, and developmental challenges. From close range it is often more difficult to see – than from a distance, within a world-wide frame – what the most original features are in one's own educational culture: where the strongest points are, where the best potential; and on the other hand, what is weak, stagnant or problematic.

An international evaluation process – defining the basic concepts, creating a shared evaluation basis, designing the instruments, planning for the analyses, even deciding on the contents and form of reporting – is a very revealing and illuminating experience for the people involved. I can assure this from my own experience: For the IEA studies I have been involved in the Reading Literacy Study (RLS) and also in the Second Information Technology in Education Study (SITES); and for the OECD projects, in the Second International Adult Literacy Survey (SIALS) and in the Programme for International Student Assessment (PISA).

It is especially at the international work meetings that the cultural differences stand out as frequently multiplex, divergent views and heated arguments, where national or regional values, traditions and visions become manifest, as often recognized by the debaters themselves, as well. Unfortunately, these multicultural encounters usually remain undocumented. They might sometimes offer at least as valuable material for the comparisons across educational cultures as the actual evaluation results available to the public. International evaluation work provides, in my opinion, an unparalleled basis for reflection and a genuinely multicultural forum for self-reflection, as well, and moreover, for critical discussion about an instrument for national reviewing and the development of educational actions.

I cannot help taking up an example from the IEA Reading Literacy Study and the heated debate then about the fundamental nature of literacy. The discussion started out by choosing the task type (multiple choice vs. open-ended questions) for the instruments. The American, British, and Scandinavian representatives were strongly in favor of open-ended questions, while others chose a more moderate both-and policy; most of the representatives of Asian countries, both from the more and from the less developed countries preferred multiple choice questions, though for different reasons. The discussion was soon taken to a deeper level of conceptual definitions, in other words, what the term literacy really meant in the different countries, cultures and languages and what was meant by its evaluation. In the agenda, one hour was reserved for this discussion. Well, it took the whole day and part of the next one. The discussion was extended to reflection on a paradigm shift in literacy and expanded to various theoretical approaches, to linguistic, psycholinguistic, functional, even to social dimensions. The battle between different cultures was passionate but honest. It was simultaneously an academic, theoretic-conceptual debate while taking stands on educational policy on the needs and uses of information.

In all, international evaluations raise the value of national evaluation work significantly (Leimu, 1987; Linnakylä, 1995):

- They provide a broader framework – beyond any particular educational system – in which to evaluate national systems and their outcomes as well as to explore various educational cultures, learning environments and pedagogical solutions.
- They provide an opportunity to get to know different ways of solving problems in education and to assess the effectiveness of these solutions.
- They yield shared conceptual models, evaluation strategies and instruments which can serve as a basis for qualitative and quantitative comparisons of education, and create an empirical database that has continuous significance to research.
- They enable contrastive evaluation of our own educational system, including its qualitative standard and resource allocation, in order to improve the system, learning environments and teaching.
- They give an opportunity to enhance international co-operation in research and education in order to promote global responsibility for education.
- They provide an opportunity for smaller nations also to display their educational and school culture as well as educational research internationally.

A small nation has, of course, better chances of making its educational system and pedagogical solution known abroad when it reaches a high performance level in international evaluations. High achievements make other countries interested in the explanatory and underlying factors of this success. Such factors usually relate to learning environments, resources and contexts. The international interest also encompasses further studies, more detailed contrastive analyses, even observations of teaching situations. In connection with the RLS, for instance, the results made many researchers interested in the secrets of the Finns' literacy skills (Elley, 1992; Linnakylä, 1993). It led into many subsequent elaboration and comparative studies on the results (Binkley & Linnakylä, 1997; Linnakylä, Törmäkangas & Tonnessen, 1997) and critical analyses, as well (Cumming, 1996). If I may say so, some of these last-mentioned analyses seem peculiar from our national point of view and show that interpreting international findings calls for more than just superficial general knowledge; it requires a deeper understanding of the national education culture (e.g. the Swedish-speaking minority).

One of the most interesting further studies following the RLS was a large Danish case study concerning Nordic countries and based on observation of teaching and on interviews with various interest groups. The study involved comparison between the primary reading instruction given in the high-achieving countries – Sweden and Finland – and that given in Denmark, whose performance level was among the lowest. This study also involved testing the students anew (just in case!). A central result of this Danish study was the finding how significant the teachers' and parents' expectations are. Teachers' expectations seemed to be related to educational arrangements. When a class has the same class teacher throughout primary school, the teacher learns to know the pupils well, but then the expectations tend to be too low (Sommer, Lau & Mejding, 1996). – Typically enough, it also tends to be easier to find funding for extended research in the low-achieving countries, in particular. Bad news is good news for future research funding.

Intertwining International and National Interests

At their best, international evaluations are attached to qualitative case studies or action research, which cast new light on the results and, preferably, draw nationally a sharper and more colorful picture of students' achievement and of the related learning-environmental and contextual factors. These substudies are often, and particularly, closely connected to the developmental interests of education and teaching.

International evaluations also provide a useful basis for national assessments with monitoring and follow-up interests. For example, in Finland the national assessments in 1991 and in 1995 included sections by means of which changes in mathematics and science, as well as in literacy skills could be monitored at the national level, as the section contained partly the same tasks as the previous international studies.

As for national assessments, I would insist, however, that the primary emphasis should be put on national values, goals and standards with a view to drawing a more colorful profile of the learning outcomes, particularly in the disciplines and content areas not tested internationally. In the Finnish context this would mean, more specifically, assessments in foreign languages, history, arts and crafts and other practical subjects as well as skills in the new technologies. We should also do more extensive evaluations of those areas, especially, where we seem to have problems according to the international findings. In the Finnish school system such areas include self-assessment skills and self-esteem, attitudes to studying, problem-solving and logical thinking, empirical learning skills and the quality of school life, especially teacher-student relations.

International and national evaluation schemes should be intertwined so that they would complement each other without straining the schools and students too much. Presently, it seems that the upper level of the Finnish comprehensive school will be faced with several, international and national, assessments scheduled for almost the same years. These will include studies of literacy skills, mathematics and science, and possibly also school-leaving examinations. Despite sampling, this will be a burden on the schools and students, but also uneconomical and senseless otherwise. It would be wiser to attach the international studies to complementary and elaborate national options, and to save separate assessments for those areas that are not covered, or prove problematic in the international evaluations.

National Challenges to Future Evaluations

Evaluation targets of national – and perhaps also of international – interest include areas related to the changes in the educational system and learning environments, such as the effects of decentralization on educational equality, efficiency, excellence, emancipation and empowerment among regions, municipalities, schools, and students. The change of teacher expertise in the new situation would need to be studied, as well. Of Nordic interest are also the consequences of lowering the school-starting age, in terms of equality, learning, self-esteem and school satisfaction. An interesting curricular issue is subject-specific versus integrated instruction.

In particular, new challenges to national, and – in my view – also to international evaluation are presented by the role of the new technologies in opening up the educational system and learning environments. It is not merely a question of changes in resources, materials or methods but of opening up the whole context of education, the pedagogic culture and the learning environment. Today, almost all Finnish schools have some co-operation with a foreign school and Finnish children already work on joint projects for example in science (e.g. the Globe project) together with their American, German, Italian, Japanese, or Swedish peers. They compile and produce information together, comparing and analyzing it, and report the results to each other and to others via the information networks. Such network-supported learning environments force us to broaden our views on learning and traditional subject-specific skills and knowledge. The cognitive and individualistic concept of learning has already become limited. Learning is now increasingly attributed to experiencing, where feelings and the fascination of doing are strongly involved. There is also strong reliance on interaction and collaboration. Learning transcends the boundaries between time and place, knowledge and experience, instruction and entertainment, history and future. However, free access to and a free flow of information lead us now more often to face ethical and moral issues and considerations. This area is one of the major challenges facing both national and international evaluation in the near future and needs to be taken seriously, not only in a school questionnaire but also when designing the framework for and defining the main concepts of the future evaluation studies, regardless of the subject areas concerned.

An innovative approach is needed also in international system level evaluation. If this aspect is ignored or neglected, the evaluations will soon be blamed for hindering the development of education and learning opportunities, as we all know the power of evaluation. On the other hand, evaluation – even system level evaluation – can through its power also advance this development by being innovative, bold, future-oriented. The time for reproductive studies is past. There are promising signs of new initiatives content-wise: especially the CCC (cross-curricular competencies) project by the OECD in Network A, which has been guided by the question: "What are the skills young adults need after their initial education and training, to be able to play a constructive role as citizens in society?" Recent efforts to integrate cross-curricular competencies into subject areas – reading, mathematics and science – are most challenging.

Furthermore, we should also venture into assessing learning that takes place outside school. Along with the strategy of lifelong learning, in recent years more attention has been paid to the demands arising from the world outside school, i.e. from adult and working life, in terms of the targets of evaluations and in defining school learning. While trust in transfer has been fading in the research on cognitive learning, the notion of authenticity, being true-to-life, and applicable skills have gained in significance. This presents interesting challenges to the definition of evaluation targets, both nationally and internationally. The key competencies (see Table 1) which our society will require in the near future for working life, active citizenship and continuous learning appear fairly similar, be they defined nationally or internationally.

Table 1
Key Competencies

Key competencies	
A barometer of future education (Kaivo-oja, Kuusi & Koski, 1997)	**OECD / Lifelong education (Cochinaux & de Woot, 1995)**
	Reading and writing literacy, numeracy
Intellectual flexibility in changes	Creative and critical thinking
Getting along in a foreign culture and accepting dissimilarity	Knowing different cultures and relating them to one's own culture
Use of electronic communication systems	User skills for information technology systems
Knowledge of and competence in mathematics, science and technology	Versatile competence in mathematics and science
English and another foreign language (French, German, Chinese)	Proficiency in a foreign language
Social skills and skills for team learning	Individual's rights and responsibilities in society
Responsibility for global problems and environmental issues	Understanding of and responsibility for the balance of nature
Ability to review science critically	
Physical fitness	Basic economic knowledge
Ability to make aesthetic handicrafts	

Both of the lists in Table 1 emphasize, at least, creative and critical thinking, familiarity with different cultures, information technological skills, an understanding of and responsibility for environmental issues and the balance of nature, competence in mathematics and science and foreign language proficiency. The international list attaches more weight on the fundamental literacy skills than the national list. They are hardly underestimated nationally, either, rather they are taken as self-evident. The national list, on the other hand, emphasizes intellectual flexibility in changes, foreign languages, communication skills, physical fitness as well as arts and crafts. Although such lists are only suggestive guidelines, they may still highlight some viewpoints worth bearing in mind when planning evaluations with respect to working life and the future.

Innovativeness should not be restricted to the selection of evaluation targets, however. It should also be manifest in the methodological approach as richness, cultural pluralism and boldness. There are some promising initiatives to this effect: One was a project based on video recordings of teaching situations and carried out in connection with the IEA/TIMSS. Another example is the IEA/SITES with its three-module structure, which includes both quantitatively oriented subprojects and qualitative case studies in order to find out innovative, school-based pedagogical solutions.

BEST COPY AVAILABLE

We should also break out of the chains of psychometry, although we need to be aware of the statistical conditions and seek new possibilities for the methods, especially for analyses of large data sets. This concerns, for example, scale construction with multiple parameters, multilevel modeling in explanatory analyses and profiling of various subgroups, e.g. by exploiting neural networks.

National Wishes Concerning Future Evaluation and Research Strategies

Both the IEA and OECD assessment frameworks and instruments have been criticized for cultural bias or even for colonialism (Cumming, 1996). The criticism has been strongest, however, among people who have not been involved in these evaluation efforts, being thus unaware of those multicultural discussions which I described earlier, and which always take place when defining the basic concepts, when constructing the evaluation framework, and when designing the instruments. Yet, it must be admitted that in the IEA studies, for instance, the social, cultural, or educational characteristics of the developing countries have not been taken sufficiently into account. Within the OECD, perhaps, the cultural and developmental differences between countries are not quite as big, and therefore comparisons between the member countries with respect to learning achievements may be fairer. On the other hand, within the OECD, in the International Adult Literacy Survey, for instance, various cultural and linguistic problems can reach an acute stage, as well. Coming from a small country with a strange language, one would hope, of course, that all participating countries could take part in the planning and implementation on an equal basis. The problems in this respect have been largely due to our own national policies, in fact, as confirming the funding may have taken such a long time, that the projects had been up and running long before we were finally able to join in. I hope that national governments and ministries take their responsibility for educational evaluation more seriously now that certain obligations are understood to be national obligations rather than enterprises of individual research institutes.

International evaluation projects are, at their best, joint ventures: planned, implemented and reported together. In the national context this has not always been understood, rather – due to our joining in so late – it has been assumed that the theoretical and conceptual framework and the instruments come from somewhere as given, and the national contribution is restricted to carrying out the instructions received, which from the national viewpoint feels neither very inspiring culturally nor very challenging as a task. In the International Adult Literacy Survey we were able to join in only in its second phase, so even the proficiency scales were set and ready, based on the first cycle of the evaluation. We had no influence on the theoretical and conceptual framework, nor on the instrument design. Naturally, we can still bring in our contribution at the stage of further analyses. It would be desirable that all Nordic countries, for example, continued their analyses and thus intensified the Nordic dimension also through an investigative approach.

Reporting has been a stage where the participating countries have usually had rather limited possibilities of contributing. The reporting and publication rights and responsibilities have been vested in the international coordination center and in the executive board. This has been the case both in the IEA and in the OECD projects. Even if these coordinating and executive bodies keep to themselves the editorial and publishing rights of the first reports, usually there are still plenty of material and an abundance of questions with education policy, research or pedagogical interests awaiting further analyses and reporting by a wider range of editors, as well. Both internationally and nationally, the data would often call for more thorough analysis and reporting. Especially, one would wish for joint ventures on further studies by countries that share cultural or regional interests. In the IEA Reading Literacy Study we managed to do this to some degree but not without protests.

The international databases should be made available to all participating countries as soon as possible, so that the national centers could do further analyses of and extensive research on the data. Indeed, this is what is often promised at the beginning of a project, but more rarely have these promises been fulfilled later. For instance, our institute has sent many letters with official requests to get the IEA/RLS international data for further analysis. Needless to say, we are still waiting. I really hope this practice would become more open and flexible, and above all, speed up a little.

The databases and further analyses would be invaluable both in research and in training new evaluation experts. In fact, it is desirable that researcher training would be included as one component in national and international co-operation schemes. This issue deserves our earnest attention in the years to come. I know that in many countries it is hard to find and recruit new, qualified evaluation researchers these days. This is certainly true for Finland. As a result of the recent interest in qualitative research on schooling and education – which, of course, is valuable as such – methodological competence among young researchers has become one-sided, and they have little knowledge of the newest statistical methods and computer software, which are necessary when dealing with large data sets. By the same token, we should strengthen their knowledge of evaluation theories, processes and implementation strategies, and also their competence to deal with social changes, especially with the economic and cultural ones. If we are going to meet the obligations of international evaluations in the near future, we need to pay serious attention to the further training of evaluation experts and researchers. In this work I would like to see international co-operation as a source of enriching support. Large international projects offer, as such, unique educational events and experiences to everyone involved. But we also need new people in the area, and there are certainly opportunities for their further training in these contexts. The resources just need to be taken into wider use.

Naturally, the training of evaluation researchers must be taken seriously at the national level, as well. Evaluation has not been considered real science, and therefore corresponding research training has not been very highly regarded at university faculties and departments of education. Now, however, eclectic and interdisciplinary approaches and social relevance are gaining appreciation, which gives reason to hope that also the Ministry of Education will see the national interests in this perspective.

References

Binkley, M., & Linnakylä, P. (1997). Teaching reading in the United States and Finland. In M. Binkley, K. Rust, & T. Williams (Eds.), *Reading literacy in an international perspective*. Washington, DC: U.S. Department of Education, National Center for Education Statistics, 139-177.

Clarkson, M.-L. (Ed.). (1994). *Suomalaiset ja koulu. Asenteita, odotuksia ja käsityksiä. OECD:n koulutusindikaattoriprojektin D-Networkin Suomen tutkimus* [The Finns and school. Attitudes, expectations and views. The Finnish study pertaining to the OECD's educational indicators project Network D]. Helsinki: Ministry of Education & National Board of Education.

Cochinaux, P., & de Woot, P. (1995). *Moving towards a learning society: A CRE-ERT report on European education*. Geneva: CRE & Brussels: ERT.

Cumming, J. J. (1996). The IEA studies of reading and writing literacy: A 1996 perspective. *Assessment in Education, 3* (2), 161-178.

Elley, W. B. (1992). *How in the world do students read?* The Hague: IEA.

Guba, E. G., & Lincoln, Y. S. (1989). *Fourth generation evaluation*. Newbury Park, CA: Sage Publications.

House, E. R. (1990). Trends in evaluation. *Educational Researcher, 19* (3), 24-27.

House, E. R. (1995). Evaluoinnin futuurin perfekti [The future perfect of evaluation]. In S. Takala (Ed.) *Arviointi ja koulutuksen laadun kehittäminen* [Evaluation and the enhancement of the quality of education] (pp. 23-32). Jyväskylä: University of Jyväskylä, Institute for Educational Research.

Kaivo-oja, J., Kuusi, O., & Koski, J. T. (1997). *Sivistyksen tulevaisuusbarometri 1997. Tietoyhteiskunta ja elinikäinen oppiminen tulevaisuuden haasteina* [A barometer of future education 1997. The knowledge society and lifelong learning as challenges to the future] (Opetusministeriön suunnittelusihteeristön keskustelumuistioita 25). Helsinki: Ministry of Education & Turku: Finland Futures Research Centre.

Leimu, K. (1987). *IEA-toiminnan esittelyä* [Introducing IEA activities]. Unpublished manuscript, University of Jyväskylä, Institute for Educational Research.

Linnakylä, P. (1993). Exploring the secret of Finnish reading literacy achievement. *Scandinavian Journal of Educational Research, 37* (1), 63-74.

Linnakylä, P. (1995). *Lukutaidolla maailmankartalle. Kansainvälinen lukutaitotutkimus Suomessa* [Global access through reading. The IEA study of reading literacy in Finland]. Jyväskylä: University of Jyväskylä, Institute for Educational Research.

Linnakylä, P., Törmäkangas, K., & Tonnessen, F. E. (1997). What is the difference between teaching reading in Finland and Norway. In J. Frost, A. Sletmo, & F. E. Tonnessen (Eds.), *Skriften på veggen. Hva skjer med var leseferdighet? – en antologi* [Writings on the wall. What is happening to reading skills? – an anthology] (pp. 97-122). Copenhagen: Dansk Psykologisk Forlag.

Messick, S. (1992). *The interplay of evidence and consequences in the validation of performance assessments* (Research Report No. 39). Princeton: ETS.

Norris, N. (1995). Koulutusohjelmien arviointi [Evaluation of educational programs]. In S. Takala (Ed.) *Arviointi ja koulutuksen laadun kehittäminen* [Evaluation and the enhancement of the quality of education] (pp. 33-38). Jyväskylä: University of Jyväskylä, Institute for Educational Research.

Scriven, M. (1991). *Evaluation thesaurus.* Newbury Park, CA: Sage Publications.

Scriven, M. (1994), Evaluation as a discipline. *Studies in Educational Evaluation, 20* (1), 147-166.

Sommer, M., Lau, J., & Mejding, J. (1996). *Nordlaes – en nordisk undersögelse af laesefaerdigheder i 1.-3. Klasse* [Nordlaes – a Nordic study of reading skills in grades 1 to 3]. Copenhagen: Danmarks Paedagogiske Institut.

Standards for educational and psychological testing. (1985). Washington, DC: American Psychological Association.

Rainer H. Lehmann
Humboldt University
Berlin
Germany

Staking Claims for Quality in System Evaluation in Germany

Introduction

The given topic carries an ambiguity which really marks a core issue in the current and growing interest in system evaluation. What is meant by "staking claims for quality in system evaluation"?

The claims to be made could refer either to technical quality to be attained by exercises in system evaluation, or to quality as a set of system goals in education, intended to be reinforced by system evaluation studies.

At first sight, the first interpretation is the more plausible one. It would appear safe enough to insist on the highest attainable methodological standards in system evaluation – a proposition to which no one could really object. So, it may be more of a challenge to explore the second interpretation of the topic: Can one reasonably expect of large-scale surveys, which are the standard technique in system evaluation, that they will affect system goals and their interpretation by the actors in the system (such as policy makers, school principals, teachers, students, parents)? If so, which are the mechanisms through which the results of such surveys develop feedback on the system itself?

A longitudinal survey – in fact a census-type assessment program – in the city of Hamburg (which is one of the 16 states of the Federal Republic of Germany) may serve as an example on the basis of which some key aspects of the topic can be illustrated. These are the following:

- aspects of justification for system evaluation studies;
- aspects of designing analyses in system evaluation;
- aspects of reporting the overall results to the public;
- aspects of linking system level and school/class level evaluations.

Some conclusions as to expected long-term effects of evaluation on the educational system as a whole are intended to substantiate the claims made in these opening remarks.

Aspects of Justification for System Evaluation Studies

System evaluation is a rather costly exercise, and it is normally not launched without good practical reasons. It is not *just* done in order to present a balance of achievement (or learning outcomes) in an educational system. Rather, the driving force behind it is often some evidence for – or at least suspicion of – gaps between overarching goals in society and the actual state of affairs.

The early studies centering around the concern for minimum educational achievement levels, in particular literacy levels considered essential for economic development and military strength (!), are probably interesting examples to be mentioned here (e.g. Rice, 1913). In hindsight, they may perhaps better be regarded as precursors to the paradigm of system evaluation, because little, if any, attention was paid here to the structure and constraints of the system as such. There are, however, contemporary parallels which shed more positive light on this approach. The participation of low-income countries in international comparisons as conducted by the IEA (the International Association for the Evaluation of Educational Achievement) may not seem justified, if only country rankings are considered. Yet, it does appear quite rational under the premise of forced development and on the basis of criterion-referenced interpretation of the achievement scales used. Such a perspective links the outcomes of the evaluation study back to the issue of attaining overarching national goals under constraints of available resources – quite a modern approach as will be seen shortly.

From a methodological point of view, system evaluation is much indebted to the large studies, such as the Coleman report (Coleman et al., 1966) and the Plowden report (Department of Education and Science, 1967), which were guided by the concern for more social equity to be facilitated through education. This becomes particularly evident when focusing on countries such as Finland, Sweden, or Germany, where an early connection with the issue of selective versus comprehensive schooling was established – with all the implications of controlling and paying tribute to the differential entry characteristics of the students involved (cf. Fend, 1982). Again, there is the clear notion of overarching societal concern and there are undeniable linkages between evaluation criteria, the methodology of an evaluation study, and the definition of "quality schooling" in the subsequent public discourse: This line of research derived its justification from sufficient initial evidence for the mediating or even reinforcing role of the educational system in the production and reproduction of social inequality. Inasmuch as it confirmed this evidence, it helped to change priorities of policy, organizational structures, and, perhaps most importantly, the perceptions and perspectives of teachers acting in the field.

For several reasons, the focus of attention in system evaluation has undergone a noticeable change since, with subject-matter achievement and, to some extent, subject-transcending ("cross-curricular") competencies once again given high priority. One of the reasons is attached to the growing awareness of global competition, with a strong view of education as a source of productivity. As is demonstrated by the International Adult Literacy Survey (OECD & Statistics Canada, 1995), for instance, the Human Capital Paradigm (cf. Becker, 1975) at last begins to have effects on public discourse, after a long period of relative neglect. Secondly, the budgetary constraints on educational expenditure which are making themselves felt in a number of countries, highlight the notions of added value in education and of accountability for spending public funds. Also, the very term "education" is being redefined: Decreasingly, it refers to an obligation of the individual to fulfill formal requirements as made by the state (the traditional version) or the obligation of the individual to perfect himself or herself (the 'enlightened' version); increasingly, it is perceived as a right to a public service. This leads back to the pivotal role of the concept of accountability. In this context, system evaluation appears not only 'justified'; it is inevitable if the stability and the legitimacy of the social and political system itself are to be maintained.

The Hamburg study (cf. Lehmann & Peek, 1997), which was mentioned above, illustrates this point quite aptly. It was initiated at a point of time (1994) when several of these lines of justification converged: Firstly, there was some indication from a test calibration exercise in the field of spelling that primary school children in Hamburg were attaining substantially lower performance levels than aspired and possibly also lower than elsewhere in Germany (May, 1994). Secondly, the Ministry of Education in Hamburg came under criticism, even from within the Social Democratic Party in the Government, let alone from conservatives and business, that too little attention was being paid to educational achievement. Finally – and perhaps most interestingly – the intended (and meanwhile implemented) devolution of decision-making power to schools under the heading of "school autonomy" entailed a whole new issue: It was at least plausible to raise the question whether or not such deregulation would increase rather than minimize the between-school variance in educational opportunities and thus unintentionally open up a new source of social inequality. Under these circumstances, it was almost imperative to call for a system evaluation study in order to achieve two things: to develop a strategy of discourse which would meet the public complaints about low achievement levels and to forge an instrument which could offset the unintended consequences of educational policy decisions. It is obvious how these two elements are both related to the more fundamental issue of the legitimacy of the government. So, a decision was made which is quite radical if judged by German standards and traditions: In September 1996, all 13,000 students of grade 5 (the beginning of lower secondary school) were obliged to participate in a test program covering language comprehension, reading comprehension, spelling, document literacy, and mathematics. At about the same time, it was determined that these students would be tested again in 1998 and 2000, in order to assess the growth within classes, schools, and educational programs (tracks), taking possible effects of social selection into account.

Aspects of Designing Analyses in System Evaluation

If an exercise such as this one is to have political effects, its design must satisfy at least two conditions: Firstly, the evaluation criteria must be closely related to public concerns, and secondly, the quality of the subsequent analyses must be such that any conclusions drawn do not immediately crumble under criticism from the scientific community (or other expert groups).

How much effort is required to define the criteria in such a way as to make the results of a study acceptable, convincing, and even compelling in the eyes of the public is well illustrated by the Third International Mathematics and Science Study (TIMSS). In particular, the ingeniously designed and laborious Test-Curriculum Matching Analysis (TCMA) of TIMSS is an excellent example of dodging criticism based on alleged cultural and curricular specifics of the participating countries. Conversely, the system evaluation studies focusing on comprehensive schooling have been followed by controversies centering around the choice of evaluation criteria. This case is particularly interesting because it is always possible to escape the consequences drawn from an evaluation study by declaring its underlying criteria irrelevant. In terms of discourse structure, this is the 'last line of defense': If somcone's practical aims and/or theoretical beliefs are threatened by the results of an evaluation study and if no technical faults can be identified, the last resort is to challenge the criteria built into this study. Incidentally, this could also be observed in Hamburg when some interested groups attempted to base their defense on criteria impossible (or, at least, very difficult) to measure, such as "creativity" or the "ability to work in a team", and to declare these as the genuine core of teaching and learning in schools.

The second element mentioned, the technical quality of the evaluation study itself, concerns the theoretical and methodological expertise of the evaluator. It may be interesting to note here that the early studies of the IEA were much more theory and methodology driven than some of the later ones. In his introduction to the report on the First International Mathematics Study, for instance, Torsten Husén (1967) developed the key concept of the productivity of an educational system, which was later elaborated into the "Educational Productivity Model" by Herbert Walberg and his group (Walberg, Pascarella, Haertel, Junker & Boulanger, 1982). Similarly, parts of the Six Subjects Survey were closely linked conceptually with John Carroll's "Model of School Learning" (1963), a line of thinking which did not get much attention in the later studies.

Judging from what was said above, the notion of educational productivity can be expected to be close to the center of concern in present-day activities in the field of system evaluation. It is focused on the relationships between the input and output of the system and tries to model the mediating processes in such a way as to facilitate interventions which maximize the output under given external constraints. Theoretically, this mode of thinking converges with recent fields of study such as research on educational effectiveness (cf. Scheerens & Bosker, 1997), new models of educational management based on economic theory (cf. Dubs, 1996), and theories of innovation (cf. Fullan, 1991). Practically, it is highly compatible with the more general concept of accountability in democratic societies. It is perhaps this notion which is particularly important in that it transcends the dimension of more technocratic system management in the direction of socially responsible decision-making.

Rainer H. Lehmann

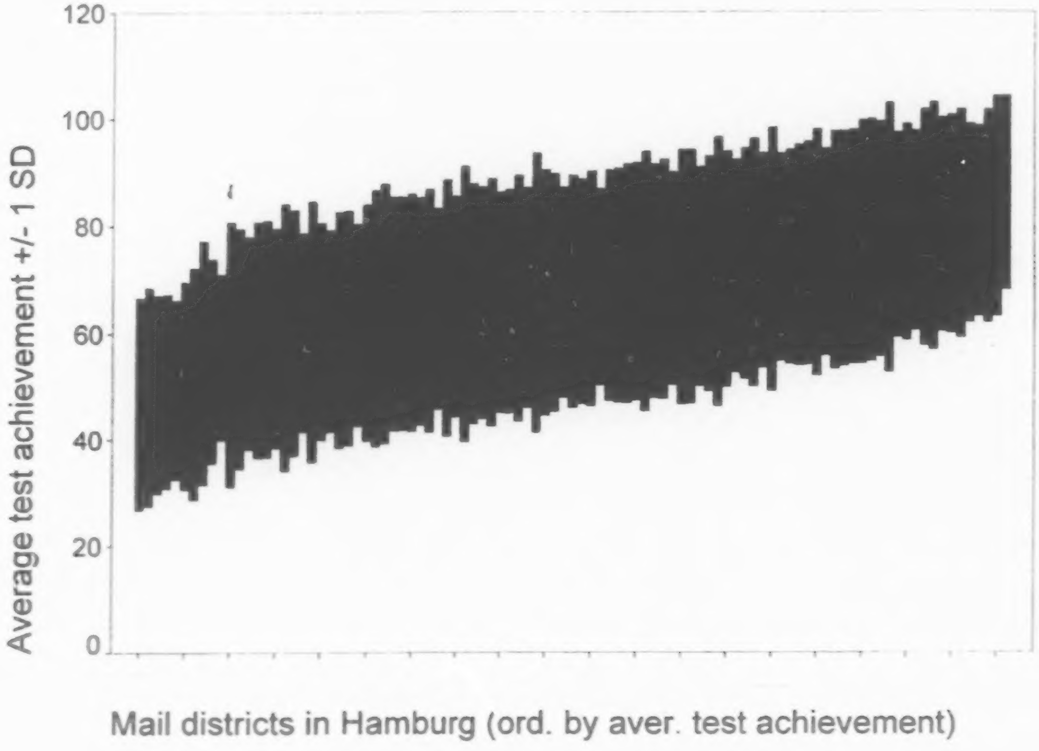
Aspects of Reporting Evaluation Results to the Public

System evaluation studies are, at least in most cases, conducted under contract to some government or public agency, and thus the results are first reported to that institution and then passed on to the public. It cannot always be taken for granted that the contract-awarding agency will be in a position to fully appreciate the potential as well as the limitations of the study. This will be much less the case as regards the general public receiving press releases and political statements based on the scientific report. So it is, perhaps, worth devoting some attention to this area which is crucial in establishing the link between evaluation research and educational policy.

It may be worth noting that, probably due to such considerations, the nature of evaluation reports changed considerably in the 1990s, moving away from statistical detail in the direction of easy-to-grasp everyday concepts such as plain percentages, bar charts and the like. This is particularly true for North America with its pervasive evaluation culture. Whatever the statistical sophistication behind (usually relegated to lengthy technical reports), the message to be conferred must be clear, straightforward and, if at all possible, visualized.

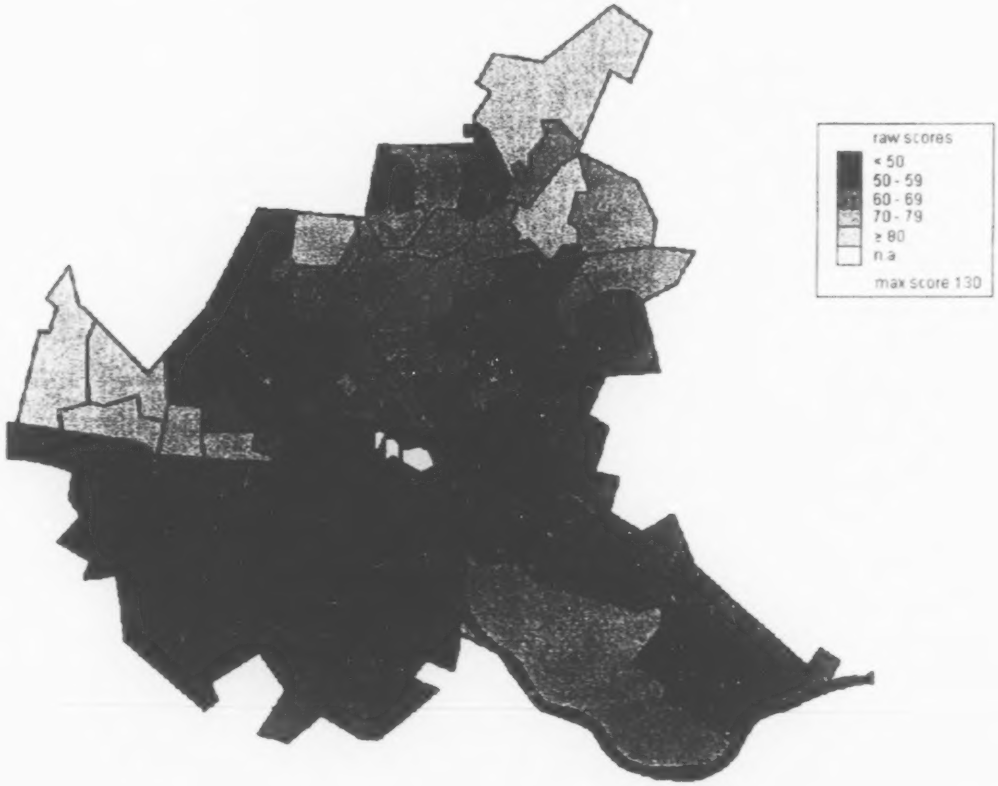
Two instances from the Hamburg study can, perhaps, serve as examples. Several questions in the catalogue of research tasks issued by the Hamburg Ministry of Education referred to the heterogeneity of learning outcomes at the end of grade 4, i.e. the end of primary schooling, because it was suspected that socio-economic and socio-cultural background factors associated with the population structure within city districts would have a sizeable impact on mean achievement levels in schools and classrooms. So, the tested population was divided into geographical zones according to the zip code in the address of the primary schools, which recruit their students almost exclusively on the basis of geographical principles (in order to minimize distances from the students' homes). As it turned out, this expectation was quite justified: Approximately 87% of the variance between zip code areas could be accounted for by aggregate factors such as the mean level of education of parents, the percentage of immigrants, or the unemployment rate in the area. Figure 1 attempts to visualize these findings.

Figure 1. The test achievement of students from grade 5 in Hamburg by neighborhood (mail districts).



The bars in the graphical representation signify the mean achievement level in an area ± 1 standard deviation. As far as could be seen from public reactions, the message of this chart was received as intended: It was meant to serve and to be understood as a sign that it is quite unacceptable that top-performing students in one area hardly reach the level of the lowest-achieving students in another. At the same time, it was an 'honest' representation of reality in that it did not conceal the fact that only about 10% of the total between-student variance is associated with the social structure of the neighborhood. For this reason, a graph of this type is probably to be preferred to a second, much more suggestive mode of presenting the same findings, this time in the form of a map (Figure 2).

Figure 2. A map of the city of Hamburg. The mean test achievement (raw scores) by mail district.



The similarity of this map with corresponding maps describing the regional income distribution and other social indicators in a published "Social Atlas of Hamburg" (Podszuweit, Schütte & Swierkta, 1992) is striking, and also its coincidence with common perceptions of the attractiveness of neighborhoods. Nevertheless, the very fact that in this chart the 'variance-within-areas' is omitted subjects it much more to simplifications and to the mere reproduction of prejudices. Admittedly, a map of educational disparities was also published, but only *after* the one mentioned above and accompanied by due explanations of the differences between the respective variance components.

In any event, the need to use easily comprehensible numerical and graphical representations and lay language in presenting evaluation results does put evaluation

BEST COPY AVAILABLE

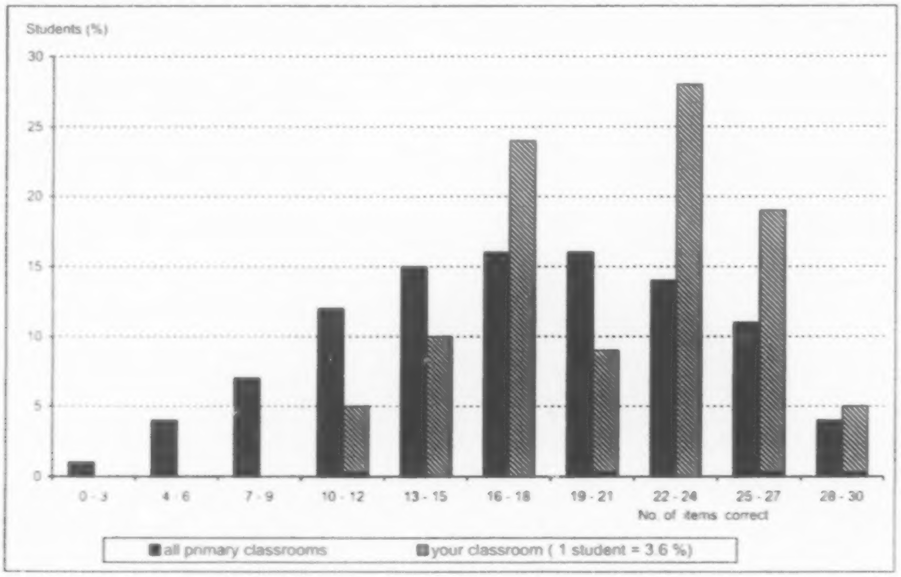
researchers in a dilemma, only partially resolved by the separation between reports geared to the general public on the one hand and technical reports as well as scholarly works on the other. In the present case, for instance, probabilistic scaling techniques were abandoned in favor of raw scores and percent-correct, because it was felt that the public, including the teachers, would not be met if sophisticated scaling techniques were used. Inevitably, though, this problem will rise again as the study moves along its longitudinal design: Achievement growth, measured at two-year intervals, can *only* be demonstrated on the basis of bridging techniques facilitated by Item Response Theory. Then, at the latest, the use of abstract achievement measures can no longer be avoided.

Aspects of Linking System Level and School/Class Level Evaluations

The problem of feeding evaluation results into public discourse and, in particular, back to the teachers concerned leads to the question of how such results should be presented in order to be conducive to the quality of schooling. This problem is particularly acute in a context where school development is primarily perceived as a prerogative of the local school faculty, and it is exacerbated by the popular catchword "school autonomy", which has found its way into school legislation in several German states as well as in other countries.

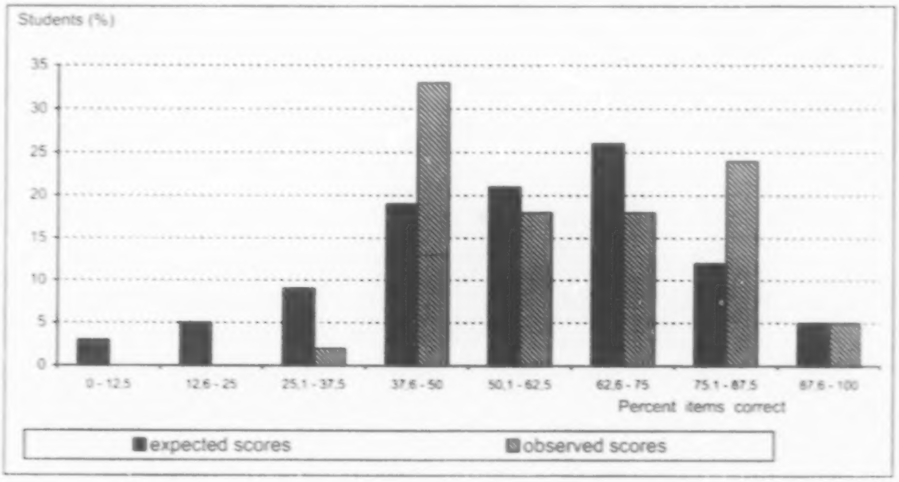
It is probably a matter of ethics to inform teachers about the achievement levels attained in the evaluation by their classes. However, even if this feedback is given in criterion-referred terms (which once more points to the necessity of using modern probabilistic test theory), this information is probably insufficient to be of any real concern to the teachers – and possibly also to the parents. A frame of reference is clearly needed in order to make the information meaningful. At least, comparisons with the performance of an accepted reference group (such as all students in the tested cohort) must be possible (Figure 3).

Figure 3. The feedback graph for an individual classroom – subscale Mathematics.



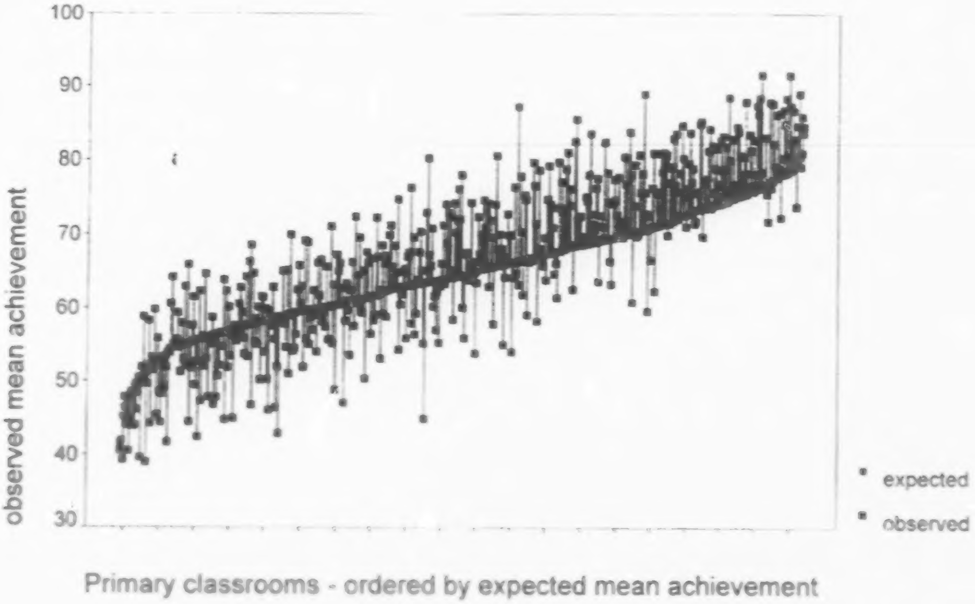
Better still, and much more in line with the paradigm of criterion-referenced testing, is a comparison with what was to be expected under the known input characteristics of the learning group (Figure 4).

Figure 4. The feedback graph for an individual classroom – the distribution of the observed total test scores vs. the distribution of the expected scores, conditioned by student characteristics



The technique used for this purpose is basically a regression-based algorithm of imputing scores, which has proved its value in North American large-score assessments as well as in TIMSS (cf. Rubin, 1987). For the German context, however, the notion of being compared to an 'expected value', which fends off criticisms challenging the fairness of the exercise, is quite new. With this information in hand, school faculties can hardly avoid beginning to reflect on possible causes of discrepancies between observed and expected results. The same holds true for the agents at the system level: If it is shown that such discrepancies exist, they cannot easily escape the conclusion that measures have to be taken to reduce the gaps between expected performance levels on the whole and, in particular, those differences where the observed level is lower than the one expected (Figure 5).

Figure 5. The mean test achievement of recombined primary classrooms (end of grade 4) – the observed vs. the expected class means, conditioned by student characteristics.



Thus, at both levels the agents are strongly urged to subscribe to and agree at least partially on the criteria underlying the evaluation. While this is not a surprise in the case of the policy makers who have agreed to the instrumentation of the study early in the process, it is all the more important in the case of the teachers. The response encountered in Hamburg so far, however, is by and large very encouraging.

Conclusion

It was attempted here to demonstrate that there is a rather close logical connection between the employment of evaluation studies, based on large-scale assessment, and the quest for "quality education", even "educational excellence". One could say that one is the logical prerequisite of the other: Evaluation is a meaningless exercise if there is no definition of "quality education", and "quality education" is an empty claim if it cannot be substantiated by satisfactory evaluation results, including findings from large-scale assessments. Almost as a side-effect, it becomes visible that this relationship can only be established on the basis of ever-rising technical standards which must be met by evaluation studies today. Thus, the call for high methodological quality in evaluation finally also appears quite an acceptable interpretation of the topic chosen.

References

Becker, G. S. (1975). *Human capital: A theoretical and empirical analysis, with special reference to education* (2nd ed.). New York: National Bureau of Economic Research.

Carroll, J. (1963). A model of school learning. *Teachers College Record, 64*, 723-733.

Coleman, J. S., Campbell, E., Mood, A., Weinfeld, E., Hobson, D., York, R., & McPartland, J. (1966). Equality of educational opportunities. Salem, NH: Ayer.

Department of Education and Science. (1967). *Children and their primary schools: A report of the Central Advisory Council for Education (England)* (2 vols.). London: Her Majesty's Stationery Office.

Dubs, R. (1996). *Schule, Schulentwicklung und new public management.* St. Gallen: Universität St. Gallen, Institut für Wirtschaftspädagogik.

Fend, H. (1982). *Gesamtschule im Vergleich. Bilanz der Ergebnisse des Gesamtschulversuchs.* Weinheim und Basel: Beltz Verlag.

Fullan, M. (1991). *The new meaning of educational change.* New York: Teachers College Press.

Husén, T. (Ed.). (1967). *International study of achievement in mathematics: A comparison of twelve countries* (2 vols.). Stockholm: Almqvist & Wiksell & New York: Wiley & Sons.

Lehmann, R. H., & Peek, R. (1997). *Aspekte der Lernausgangslage von Schülerinnen und Schülern der fünften Klassen an Hamburger Schulen. Bericht über die Untersuchung im September 1996*. Hamburg: Behörde für Schule, Jugend und Berufsbildung.

May, P. (1994). *Rechtschreibfähigkeit und Unterricht. Ergebnisse der Voruntersuchung zum Projekt Lesen und Schreiben für alle*. Hamburg: Behörde für Schule, Jugend und Berufsbildung.

OECD & Statistics Canada. (1995). *Literacy, economy and society: Results of the First International Adult Literacy Survey*. Paris & Ottawa: Authors

Podszuweit, U., Schütte, W., & Swierkta, N. (1992). *Datenhandbuch Hamburg. Analysen, Karten und Tabellen zur sozialräumlichen Entwicklung*. Hamburg: Hamburger Verein für Sozialpädagogik.

Rice, J. M. (1913). *Scientific management in education*. New York: Hinds, Noble & Eldredge.

Rubin, D. R. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley & Sons.

Scheerens, J., & Bosker, R. J. (1997). *The foundations of educational effectiveness*. Oxford: Pergamon Press.

Walberg, H. J., Pascarella, E., Haertel, G. D., Junker, L. K., & Boulanger, F. D. (1982). Probing a model of educational productivity in high school science with National Assessment samples. *Journal of Educational Psychology, 74* (3), 295-307.

THEME 3

A NATIONAL STRATEGY OF EVALUATION

Vilho Hirvi
Ministry of Education
Finland

Challenges to a National Strategy of Evaluation: Visions and Expectations

The priorities defined by Finland as factors contributing to its success in international competition are a knowledge-based society and the Finnish lifelong learning strategies. It goes without saying that education, training, research, and development are the keys to the national implementation of these strategies.

Over the nineties, Finland made some major education policy choices and reforms. Some of the most important are the following:

- We are taking measures to intensify education-work relationships at all educational levels.
- We created a non-university sector to exist alongside the traditional universities.
- We have a program for increasing research input (the target was to raise the R&D funding to 2.9% of the GDP by 1999).
- We are taking joint national measures in order to raise the standard of mathematics and science instruction.
- We are taking concrete action to translate the principle of lifelong learning into practice.
- We carried through an overall reform of educational legislation (I had the pleasure of chairing the parliamentary committee preparing this reform).

Evaluation plays an important part in all these initiatives.

The new legislation reduces the former detailed and centralized regulation and delegates more discretion to schools and the organizations running them. This is why it contains fewer acts and stipulations than the previous legislation.

The objectives of education are defined in national curricular guidelines issued to schools. In addition to this, the legislation stipulates pupils' rights and duties, teachers' qualifications and the overall principles of state financing. It is clear that since schools use different means to achieve the goals, the role of evaluation as a tool in regulation is growing. Therefore, the new legislation includes specific provisions concerning evaluation.

As Finland has decided to pursue national education policy, we naturally need national principles for evaluation. I was appointed to chair a committee which drafted the National Evaluation Strategy. This committee was composed of representatives of the Ministry of Education, the National Board of Education, the Academy of Finland, the Institute for Educational Research, and the Association of Finnish Local Authorities. It is expected that this broad-based representation helps to get wide approval for the principles defined in the National Evaluation Strategy.

Our national evaluation strategy seeks answers to three questions: Why, what, and how should we evaluate?

Why Should We Evaluate?

Educational systems are changing rapidly in all OECD countries. The challenges which the knowledge society presents to education, the diversification of structures and forms, the decentralization of administration, the enhancement of internationalization, the quantitative expansion of higher education, and constraints on funding and resources are examples of these extensive changes. Education is becoming more like a customer-driven system, steered according to the goals set by each provider of education. One central trend at all levels of education is to intensify contacts between educational institutions and the world of work. This requires up-to-date information about school-work relationships.

The idea of learning as seen from a lifelong perspective makes many new demands on the implementation of evaluations. Lifelong learning with a view to personal and professional development, career change, transferable skills and matching supply and demand is essential now and even more so in the future. At the same time, more autonomy has been given to educational institutions and to those who run them. The need to know what is really happening at educational institutions and what kind of results have been achieved is increasing. Though local authorities and the educational institutions they are running are responsible for developing their education and training provisions at least partly according to local needs and circumstances, we need information on results obtained in the whole country: How is equality achieved? How is sustainable human development guaranteed? How effective and efficient is education in different parts of the country?

The number of educational evaluations grew in all European countries during the 90s. Often, the official reason is an effort to enhance the level of education. In several countries, administrators are also willing to create some kinds of inspectorate systems to control the level of standards of different parts of education. Performance-based resource allocation is also under discussion in many countries, especially as far as higher education is concerned.

The aim of national evaluation is

1. to support local education provision and the development of educational institutions as goal-oriented and open units; and
2. to produce and provide comprehensive, up-to-date, and reliable information about the prerequisites, performance, results, and impact of the educational system within national and international frameworks.

National information produced by evaluation has also created a need for international comparisons. National and local politicians are keenly following the results of international indicators. Assessments of educational effectiveness, such as those done by the OECD, are an outstanding example of high-level international co-operation in the field of evaluation.

Thus, national and international evaluation projects should generate information on the quality, content and outcome of education and training in relation to the objectives of society, working life and the individual. In addition, the chosen education policy must be evaluated in relation to social development and changing individual aims.

What Should We Evaluate?

Evaluation means interpretative analysis of the object or activity under scrutiny with a view to determining the benefits or value it produces. In a rapidly changing society it is necessary to know what we want from education and what kind of society we are willing to develop with the help of education.

Educational evaluation is based on follow-up and research information, expert knowledge, and international comparative data. Information on institutions is also needed in evaluating the system as a whole. The need for information can also be satisfied, to a large extent, through more efficient use of existing sources, such as a national examination database – its use in the evaluation of general upper secondary education should be developed further.

National evaluation should cover the following:

- educational demand and supply, access to education and training, and student flow;
- the structure and functioning of the educational system and its constituent parts;
- the interrelationship of educational quality and resources;
- development trends in education policy and changes in the educational system;
- relationships between education, the world of work, and the rest of society;
- curricula and instruction;
- learning outcomes;
- regional usefulness;
- evaluation of the effectiveness, efficiency, and economy of education

At the level of higher education, as well as at other levels, it is important to make evaluation an integral part of institutional operations and to enhance institutions' expertise in evaluation. Institutional self-evaluations must be tailor-made according to the local needs. This is necessary because in response to increasingly differentiated demand, flexibility with regard to access to programs, the content, scope, depth, and duration of programs, the means of delivery, and examinations is needed.

Comparative evaluations are becoming more difficult in practice. In order to better serve the needs of diversification, wider and more imaginative institutional profiling is expected to take place within educational systems, hence leaving less room for the categorization of institutions.

How Should We Evaluate?

National evaluation must be an open system. The organization under review and the people working in it must be informed well in advance about the purpose, timetable, and consequences of the evaluation. Evaluation must allow for local aims, interpretations, and expectations. The staff in an organization must be given an opportunity to set forth their own views on the evaluation and its findings. Evaluation must not harm the objects of evaluation in any way or complicate the activities they undertake to achieve their goals. It is important to see to it that students are involved in the process of educational evaluation and that the sources of information used and the methods of compiling and analyzing are documented and justified.

The overall aim is to produce information which is

- reliable and comparable;
- pluralistic;
- timely;
- compiled comprehensively, professionally, and systematically;
- economical – that is, only essential data is collected;
- both qualitative and quantitative; and
- compiled using well-founded methods.

Also, any subsequent analysis should openly present all the data available and the methods used. This entails examiners having a profound knowledge of the object under scrutiny and their being governed by high ethical principles, as well as their having a thorough knowledge and a profound understanding of the social and human impacts of education.

Conclusions

Evaluation and relevant discussion based on the results and on our values play an important part in reconciling the objectives of the national education system and local objectives and in identifying and clarifying common training needs. Evaluation also forms part of the internationalization of education and training. National evaluation projects must be co-ordinated with international evaluations.

It is quite clear that national and international evaluation projects are needed, together with institutional self-evaluation. But we must find out the optimum amount of evaluation. We have to obtain enough information to develop the national education system, while at the same time trying not to disturb the working of the educational staff. As the number of different local, national, and international evaluation projects is growing fast, the need to co-ordinate evaluations and assessment is increasingly necessary. It is very frustrating for educational institutions to collect different kinds of information for different purposes if projects come one after another and those who are evaluated do not understand the reasons for the data gathering. The staffs of institutions are, even without evaluations, very hardworking and busy. If they are given extra work, they must also understand why it is necessary to do it. Evaluations are so time-consuming and expensive that we must always be clearly aware of what we are doing and why. Also, there has to be a clear plan for using the information we get.

One critical point in evaluation is the expertise of evaluators, assessors, reviewers, or whatever you want to call them. There is a great demand for the training of this important staff. Here we need also international co-operation in order to create effective in-service training for them. The development of evaluation methods is another field calling for international co-operation. I hope that the OECD could be active also in these fields.

It is important that evaluation is seen as a means of creating a better future, not only as a means of looking back through a mirror. Evaluation must give us some ideas of how students can be helped to prepare to meet the challenges of the intrinsically uncertain labor market. In addition to their professional qualifications, students require a broad set of attributes in terms of personal and transferable skills and competencies in order to increase their employability in the knowledge society. Evaluation must give ideas for needs of this kind, as well.

Pentti Takala
National Board of Education
Finland

How to Use Evaluation Findings

Levels of Evaluation

There are many different levels of evaluation, ranging from the assessment of students to international comparisons of school systems and educational attainment.

When it comes to student evaluation, the main problem is variation in the criteria teachers use when assessing students. This makes it difficult to compare grades from school to school. We can see this clearly in the competition for access to further and higher education.

Self-evaluation in schools is the quickest way of improving working practices in the direction wanted, as long as this evaluation is sufficiently honest. So far the main problem has been inexperience in actually doing self-assessments.

National evaluation is the task of the National Board of Education. While the general aim of national evaluations is to serve schools, administrators, teachers, pupils and parents, and also policy makers, we aim to identify strengths and weaknesses in our school system and areas needing change. By doing this, we can manage the national framework curricula and the guidance given to schools in the best possible way – at least we hope so.

Since 1994 we have done a considerable number of in-depth analyses of evaluation. Their aim has been to get a general picture of the school system and to highlight problem areas at the national level. In this presentation I shall concentrate on projects in general (rather than in vocational or higher) education. In particular, I am going to focus on the project for the National Evaluation of Upper Secondary School Education 1994 and its impact on school development.

Almost all our projects have a similar strategy. Discussions with decision makers in the Ministry of Education give us a starting point for the evaluation. In some cases evaluation starts of their initiative alone. Later, the National Board of Education prepares a more concrete evaluation plan which defines the objects of evaluation, estimates the physical resources and staffing required and which hopefully asks the right questions. The Ministry provides the resources needed.

We also try to draw on the best expertise country-wide. We enlist the help of researchers and universities, and I am pleased to say they have become a natural part of the projects. As we want to submit the results to public scrutiny so that any shortcomings and distortions could be revealed and important issues addressed publicly, the project findings are then published and made available for comment. The ensuing public debate in the mass media is also analyzed.

Another function of this public debate is to give schools additional feedback and to generate interest in the report itself. Individual schools can reflect on their own situation in the light of the report. We regard it as very important that schools and their teachers make use of these results, and we want to help them to do this.

After the public debate – which is usually drawn-out, but to my mind essential – the National Board of Education arrives at its final conclusions and prepares a set of action points for the Ministry of Education. In our proposals, we inform the Ministry about the main problem areas and the action needed. The Ministry of Education subsequently makes political decisions, and the National Board of Education is given the task of more detailed planning and implementing the improvements. Now the work of the developers at the Board can begin.

The National Evaluation of Upper Secondary School Education

One case study that I wish to describe a little closer is the National Evaluation of Upper Secondary School Education in 1994. The focus will be on how the results of the report (Jakku-Sihvonen & Blom, 1994) have been used.

The report evaluated Finnish upper secondary schooling in 1994, starting to work it up at the beginning of 1993. Evaluation was directed at two areas in particular: educational resources and the curriculum. We looked at the resources and how they varied, learning and working atmospheres, questions of inequality and educational attainment, and the impact of upper secondary school on students' further and higher education. Research centers, such as the Institute for Educational Research at the University of Jyväskylä, and interest groups, such as the Association of Headmasters in Finland and the teachers' trade union, took part in planning the project.

The main results of the national evaluation of upper secondary schools are shown below. The list was compiled after the public debate, and it consists of 9 key points of the results.

The upper secondary school theses issued by the National Board of Education are the following:

1. Finnish upper secondary school will be of high quality in international comparison: Instruction in science will be developed to reach the international level.
2. Upper secondary school has an active role in the educational system: Skills and knowledge required for studies at the institutions of higher education and polytechnics are promoted, and the willingness to develop oneself is encouraged.
3. The matriculation examination and the upper secondary school-leaving certificate will assess knowledge: The matriculation examination and the school-leaving assessment are developed to assess skills, oral skills of languages, and specialized skills and knowledge.
4. There is a need to improve study methods: Students' autonomy in studying, individual study programs and the utilization of new technology are increased.
5. There are too few students taking courses in advanced mathematics, physics and chemistry: In-service studies for teachers are encouraged; the utilization of experimental instruction is increased and the matriculation examination is developed.
6. The objectives issued by the Council of State to diversify language studies are not reached: Studies in the German, Russian, French, and Spanish languages are increased and intensified.
7. Geographical differences in language studies: The teaching of oral language skills is intensified and international co-operation is improved.
8. There is a need to strengthen the contacts between vocational education and working life: In terms of curriculum design and the implementation of instruction, the contacts between vocational education and working life are to be improved.

9. Upper secondary school is an economic institution: Co-operation between educational institutions is promoted and the utilization of new technology is intensified.

You can see several problems (listed above) and ways of solving them in the form of statements, written up by the National Board of Education. Some of these problems were known even earlier. In fact, quite a few were, in my opinion, self-evident and publicly known, but the report made them even more apparent and convincing. (This means that we made some good guesses when targeting those evaluation areas.) The report and the discussion that followed aroused public interest and led to demands for change.

A new government started its work in the spring of 1995. While the Government's education policy closely defined the most important areas of educational provision, it recommended (in its policy program) that the standards in mathematics and science ought to be higher. This, along with the diversification and development of language teaching, became part of the education policy of that government. The practical work itself was vested in the National Board of Education, and the resources were assigned by the Ministry. This is how the Mathematics and Science Programmes and the Diversification of Language Teaching Programme started. They were the result of the evaluation process concerning upper secondary schools.

As the results for upper secondary schools reflect not only the state of upper secondary education itself, but also the state of education leading up to that level, all the projects focus on comprehensive schools, upper secondary schools and extend also to vocational schools. The National Board of Education works alongside the Ministry of Education, municipalities, schools, universities, and enterprises.

Points 1, 2, 5, and methodologically also 3 and 4 above deal with problems of mathematics and science instruction. That is why we organized – together with the Ministry – the project intended to improve the teaching of mathematics and science, which got resources and staff to concentrate on working on these problems.

The Mathematics and Science Programme

The Mathematics and Science Programme started in 1996 and will be completed in the year 2002. The aim of the program is to improve teaching in schools so as:

- to get more pupils interested in mathematics and science:
- to improve the quality of learning of different student groups;
- to encourage both boys and girls to choose courses in advanced mathematics and science in their study programs;
- to provide students in different school sectors with the skills and knowledge needed in everyday life as well as in further and higher education:
- to foster the skills students need with a view to sustainable development: their attitudes to, awareness of, and approach to work.

Some of these aims have also been used as quantitative targets. For example, we aim to have over 16,000 students taking the test in advanced mathematics (in the matriculation examination), and 9,000 taking the physics test and over 8,000 students taking the chemistry test. As for international comparisons in science education, Finland aims at being in the top quartile among OECD countries. These aims are determined by the Ministry of Education according to the Government's policy program.

Accordingly, we took some 50 pilot schools from various levels and started to co-operate with them in developing methods and trying to find out more carefully what the problems are in the classrooms at the national level. The pilot schools are supposed to develop innovative teaching models which will then be disseminated to the whole country. A team of researchers are investigating the effect of these models on teaching, so the evaluation work is going on all the time. All universities and teacher training colleges preparing teachers for the upper level of the comprehensive school and for upper secondary school have been actively involved in the program. In this way, new ideas and innovations can be adopted more easily. The financial support given to in-service training programs annually amounts to about nine million marks. Ten million marks were set aside for the pilot upper secondary schools so that they could purchase laboratory facilities, equipment and materials. In addition, school administrators (in local municipalities) have been involved in financing the activities to an extent I would not like to guess.

It is not possible as yet to predict the results of the program; it is difficult to achieve much in such a short space of time. However, this may be taken as one example of how we are using the results of evaluation.

General Aspects of Evaluation

Finally, I would like to return to some general aspects of evaluation itself. I am sure you will agree that the main aim of evaluation is to provide a basis for further development. I do not want to claim that evaluation would not have other functions, but this is certainly the most important one. An evaluation which remains in the hands of researchers and which nobody needs is, to my mind, of no use at all. Surely then, it is the evaluator's responsibility to ask the right questions which are also useful in the development work. The developers have an important task here, because they know intimately what the aims of the curriculum are.

One of the tasks of evaluation is to find out if schools have achieved their goals. Another point is that evaluation follows certain principles. These include, among other things, ensuring that the findings are reliable, with proper timing and with conclusions carefully defined, and presenting results which can be publicly criticized. It is important to find methods that give reliable answers to 'the right questions'. Methods of evaluation must be designed according to the needs of evaluation..

The third phase of any evaluation is to present a vision of development and to put this vision into practice in the light of the results. Decision makers can then identify the most essential needs for change and allocate financial resources and staff. The developers can then carry this through.

As for the National Evaluation of Upper Secondary School Education 1994, I think we were at least able to ask the relevant questions. They were asked in an appropriate way, in the right place, and they were followed by public criticism. The timing can be questioned, though, because the evaluation was published at the same time as the new *framework curriculum* and therefore could not be taken into account when writing the curriculum. It did, however, have a bearing on the implementation of the curriculum.

The results of evaluation can be used in several ways. They may contribute to the development of overall educational provision, in particular when developing curricula, teaching methods and teaching materials important at the school level. They are also useful when allocating resources and in policy making.

I have spoken about evaluation at the national level, which is a duty of the National Board of Education. One of our firm principles is that the task of national evaluation is not to produce ranking lists, but to identify problems in the educational system. Evaluation at the local and school level, called self-assessment, has the same logic. A school's assessment of itself is the first step toward improving teaching and education and ensuring quality. It is not an end in itself. The conclusions drawn from it are the most effective way of making the necessary changes, but schools need a more general framework in order to be able to compare their self-evaluations, to see where they are and where they should go.

References

Jakku-Sihvonen, R., & Blom, H. (Eds.). (1994). *Lukion tila* [The state of general upper secondary school] (Arviointi ja seuranta 5). Helsinki: National Board of Education.

Pentti Yrjölä
National Board of Education
Finland

Pillars of National Evaluations: Reconciling Research and Policy Interests in Evaluation Programs in Finland

Finnish Policy Interests

I shall be talking about policy interests from a national, rather than from an international point of view, and perhaps saying something about research work. I work at the National Board of Education, which is a national expert body in the field of the development and evaluation of education. Formerly, there were two such bodies (called Central Offices in state administration): one for general, the other for vocational education, each having a lot of administrative work. Nowadays, we no more have any administrative work. What I am trying to describe to you is the approach of the National Board of Education – how to do evaluation work especially from our point of view. The work is still in its developmental stages, so I will be very pleased to have your critical comments on it. This will be a very Finnish point of view and will have very little to do with international connections.

Finland as a nation has its particular characteristics: We only number five million people, and there are some traditional values that we respect in our society. General education, in particular, is very important, very highly valued by all social classes. We also think we are very modern and are trying to become more international all the time although you may understand that we have lots of problems especially with languages. But we will manage. We are loyal to our families, to our social reference groups and to our native country. However, during the past few years, I am afraid, these basic values have changed very rapidly because there are very profound changes going on in society, of which we have only a little real research knowledge.

For me, the educational system is a political system, based on political decisions. It is something that is very national and, of course, we believe that our system is the best for Finland. The decisions the politicians make are also the best possible decisions that can be made as far as this country is concerned. But the structural and functional logic of an educational system differs from that of industrial production and of other social services. Being a human institution, school always deals with human connections and relations, and in that regard it is very difficult to evaluate how the system is working. However, we are quite convinced that political decision makers have the right to have valid knowledge as to what purposes funds are used for, what schools are doing with taxpayers' money, and whether they are doing their best. As for the question of learning achievements, we can ask whether children are really learning the right things in school.

We have a monopoly in primary but also in university education, as we have a state monopoly on selling alcohol. The same holds true for education: The state and the municipalities have the monopoly over it, but it is a very peculiar monopoly because it is completely publicly financed. The state and the municipalities provide the funding; there are no student fees. The political values have been operationalized in political decisions, which can be found in laws, acts, or other decisions made by the Government, but also in decisions made by the National Board of Education on national framework curricula. They all are explicit, they are written up, and they must be taken seriously because they are intended to steer the work of the whole system. A special duty of the national evaluation system is to give valid information on what is going on in schools and what their outcomes are.

There is, of course, a difference between our work and research work. Researchers can have their freedom, they can even evaluate whether the values, ideas and targets are the best or adequate ones or not. But as civil servants, we are bound to political decisions. Perhaps I am simplifying this difference a little bit too much but I think that I have the right to do so because this discussion is going on in Finland.

The entire society has gone through very severe decentralization. One reason why we had such a severe economic catastrophe at the beginning of the nineties was, it was often said, that we had very large central administration and strict norm regulation nearly everywhere in society. The administrative culture which was prevalent for years was said to be wrong, and we were to change the whole procedure, open up the system and make it a more competitive one. And now we can see a totally different administrative system in Finland. It is really surprising how profound the changes have been in the last few years. The number of norms and regulations in particular has decreased. But there are still people who are missing them.

Parliament has passed a new bill for new overall legislation on the whole school system. At the national level the changes are few. But when we go to a school, we find that there are really very many contradictory developmental processes going on. At the same time, schools should be transparent, open to society, they should be flexible, and students should have more opportunities to make their own choices. Schools have the right to establish their own profiles, but at the same time they must save public money. So especially in primary and secondary education, they are made more accountable to the owners, the municipalities, as there are also tendencies toward funding based on results. These trends are very important challenges, indeed. They apply especially to evaluation: How could we evaluate the effects of such trends in school? How do they affect students' learning?

The Policy of National Evaluation

We have three different cornerstones or pillars determining what and how we are doing in national evaluations: evaluation projects, evaluations of learning achievements, and the further development of indicators.

One of our answers is to launch fairly comprehensive projects related to the school system, evaluation projects concerning a particular educational sector, or a form of schooling, or a specific theme. These projects would normally take one or two years, and the idea is to get an overall picture of what the situation is like in that type of schooling – whether they are achieving the goals and purposes that are written up in the legislation. In our view, research work plays a very important part in these projects. The method of doing project work also follows the scientific way of thinking.

After the description and analysis at the end of the project, we draw conclusions from or make value judgements of the situation. We try to answer the questions: Is the situation in this theme or in this school form reasonable or not? What are the weaknesses and the strengths? Have the national goals been achieved? Do we have any proof of added value? What about social equality? This is perhaps the most crucial point of an evaluation project.

The second pillar is traditional testing or assessments, which can be done separately or in connection with an evaluation project. We have our matriculation examination after general upper secondary school, and in the spring of 1998 we had our first national samples – a five-percent sample of the pupils at the end of comprehensive school – tested in mathematics, physics, and natural sciences after compulsory education.

In 1994 we had a very large evaluation of upper secondary education, and in 1998 we did an evaluation of machinery and electronics in upper secondary level vocational education. We also organized vocational competence testing extending over one week at the end of the spring term. The testing consisted of a theoretical part, but in 1999 it included also practical work. These student assessments, the indicators and all the research material and self-evaluation reports were used to evaluate nationally how technical schools were working, especially in the fields of machinery and electronics.

The third pillar is indicators, which I will not discuss in this presentation.

In order to unify our national evaluation work we have developed at the National Board of Education a common framework for evaluating educational outcomes. This model is used in all our activities. We use three elements in evaluating educational outcomes: the efficiency (includes internal productivity), funding (economy), and effectiveness of education. In the element of effectiveness we have learning achievements at the top, but it also includes cross-curricular competencies (CCC). There we are especially interested in learning-to-learn competencies, for which we are developing evaluation methods.

The general function of education is, of course, to have an influence on students. This influence, with its particular content, is written up in norms for us, also called state evaluators.

THEME 4

MERGING NATIONAL ASSESSMENT AND RESEARCH POLICIES

Jouni Välijärvi
Institute for Educational Research
University of Jyväskylä
Finland

Research in the Context of National Assessment

The effects of educational evaluation and assessment on the work of schools, teachers and eventually students are important, and obviously increasing (Abbott, Broadfoot, Croll, Osborn & Pollard, 1994). We may even ask ourselves whether there is already too much evaluation going on, or exactly what we are going do with all the evaluation data that is being accumulated. In this connection – and particularly when national assessments are being talked about – the question of the tail wagging the dog often comes up. In Finland, we are going through a phase where long-held ideas about developments in educational evaluation are finally beginning to be put into practice, as the next few years seem to be filling up with many national and international evaluation projects.

Introduction

As regards the question of backwash effects, the basic way of thinking in Finnish evaluation discussion may be stated as allowing the dog to take care of its own tail. National evaluation is looking for an identity of its own, especially in sample-based survey studies that examine the condition and effectiveness of basic and lower secondary education. By contrast, attempts to rank schools and students in order of quality by using national tests taken during the final phase of the comprehensive school seem to be gaining less popularity than in some other countries in Europe. Thus, the curriculum and reforms associated with it seem to retain their due importance. In this presentation, four central issues will be discussed that, in my opinion, decisively affect the role to be played by research in the context of national evaluation.

Firstly, what role will *research* be given in the implementation of national evaluation projects? At issue here is above all whether the research community should be merely carrying out specific evaluation tasks commissioned by other parties or whether they should actively contribute to the development of a system of evaluation and assessment. (House, 1990.)

Secondly, what kind of responsibilities will the research community have for the *consequences* of national evaluation? That is, how critical a role will researchers be willing and able to assume regarding the evaluation projects to be implemented? (Brown, 1997.)

Thirdly, what are the preconditions for the research community to be able to meet the increasing challenges as regards evaluation? How can we ensure the production of well-substantiated evaluation data of a high qualitative standard? (House, 1990.)

Finally, what kind of *division of labor and co-operation* between the research community and educational authorities will work best in the rapidly changing field of evaluation?

Evaluation Studies as a Part of National Evaluation

Up to the present, educational evaluation in Finland has been carried out largely in terms of research and has been dependent on the active contribution of researchers. In Finland the Institute for Educational Research (IER) has, since the late 1960s, had a major role in evaluating educational outcomes and in other national tasks concerned with educational evaluation. The IER has performed these tasks mainly in collaboration with the IEA (International Association for the Evaluation of Educational Achievement) in connection with international comparative studies of educational outcomes in different school subjects. It may be said that our picture of the (comparative) standard of Finnish education is very largely based on the findings of such international evaluation studies. Our widespread belief that we are the best readers in the world or that our mathematical and science competencies are rather low, is based mainly on the results of IEA projects.

In this sense, the effects of international evaluation studies have been and still are considerable. In many cases, a Finn finds it easier to believe something if he or she has learned of it in the context of international discussion. Educational evaluation and assessment is no exception. The consequences are not always rapid, but once things begin to change, we tend to make a point of being thorough. (See Bonnet, 1996, 1997.) This is currently demonstrated by numerous projects launched by the Government with a view to developing the teaching of mathematics and science measures justified chiefly by the results gained in IEA studies as far back as the first half of the 1980s.

Up to the last few years, researchers have been active also in starting national projects. Such evaluation studies have been carried out in co-operation with public authorities (notably with the National Board of Education), but the IER has had the chief responsibility for their actual implementation. In fact, when the Institute for Educational Research was established over 30 years ago, educational evaluation and the development of assessment methods were defined as being among its central tasks. A prominent aspect of these national evaluations has always been an endeavor to develop both the methodology and content of evaluation. This may also explain why Finland has never experienced the kind of conflict or lack of mutual understanding between the research community and school authorities that has characterized certain other countries. (See Butterfield, 1995.) For the above reasons, the research community has been allowed to take considerable liberties with values, criteria and modes of action when guiding evaluation projects. Up until the early years of the 1990s, educational evaluation and evaluation research were nearly synonymous, at least in the context of the systematic assessment of the state and effectiveness of education.

Though the relations between educational research and educational authorities are fairly unproblematic in Finland, evaluation research has often been criticized for being too far removed from current trends in education, for being bound up with hopelessly slow timetables and for producing results of little value in decision-making. However, in the 1990s the field of educational evaluation was transformed. On the national level, the National Board of Education has been an unprecedentedly active agent in evaluating education. Also, the new school legislation creates substantial obligations regarding evaluation.

The New Role of the Research Community

From an international perspective, the work of the OECD in developing educational indicators entails increasingly systematic international evaluation projects more emphatically linked with educational policies. Their timetables and the quality standards set for their implementation presuppose highly professional and long-term organization of evaluation and assessment. This situation is in many ways new. Researchers must define their role in the field of evaluation and assessment on a new basis. Our strengths are in the long and many-sided experience of implementing demanding research projects with which this interest has provided us. Further opportunities for such experiences have been dangerously threatened, though not completely eradicated, by the economic recession and cutbacks in the last few years. Competition for funds is speeding up as the volume of funding allocated to evaluation and assessment is increasing. Competition as such is a positive thing and has, at least up to now, improved rather than narrowed our scope of action.

However, it is important that we ask ourselves what kind of role the research community is willing and able to create for itself in the changing field of evaluation. Shall we content ourselves with merely carrying out evaluation projects designed by others and concentrate only on improving our competitiveness in the increasingly fierce struggle for resources? Our role may be cut down from that of actively taking the initiative in developing evaluation and assessment to that of routinely implementing regularly repeated evaluation projects. This is not what one would wish to happen. (House, 1990; Sizmur & Sainsbury, 1997.)

Many researchers have continued to play an important part, right from the early stages, also in the new national and international evaluation projects. It is highly desirable that these links survive and grow stronger. It helps to reconcile administrative and research interests and creates favorable opportunities for the successful conduct of such projects. It is in ensuring a high standard of educational evaluation that the research community has a particularly important contribution to make. Basically, the quality criteria governing educational evaluation and evaluation studies more generally have close parallels, especially as regards the practical implementation of evaluation projects. The shared goal is the production of many-sided, reliable and generalizable knowledge. We all know how crucially fulfilling these criteria depends on the careful preparation and systematic implementation of the projects. (Cuttance, 1994; Daugherty, 1995.)

New Challenges to Researchers

Another equally demanding challenge involves the constant renewal of the content of evaluation. How can assessment be made to encompass, more and more fully, the whole spectrum of instructional goals? (Bonnet, 1996; Dylan, 1996; Williams & Ryan, 2000.) A system of evaluation that is implemented mechanically and whose content remains unchanged is a threat to the continuous renewal of teaching. On the other hand, we do know what a demanding task it is to remodel the content of evaluation and assessment and extend it to new target fields. This will make such activities as developing the CCC (cross-curricular competencies) indicators in the OECD/INES project increasingly important in the future and lead to their application to more and more new sectors.

Since the work presents many challenges, and because the methodological challenges in particular are overwhelming, such work requires, both on the national and on the international level, all available expertise. As we are here involved with educational objectives that are often in themselves universal but whose manifestations are very strongly bound up with national cultures, it is also necessary to boost interaction between national and international evaluation units and individuals. At the same time, as research is an essential aspect of educational evaluation and assessment, the research community must also assume a critical role vis-à-vis evaluation. The challenge facing us is how successfully we will be able to combine these two roles. (See Eisner, 1984.) In earlier evaluation studies we always stressed the necessity of close collaboration with schools in preparing the projects and in utilizing their results. We have, among other things, attempted to ensure that the schools in the sample will get the results of their own students within a few weeks after the data has been gathered, together with comparative data needed for their interpretation.

We are moving toward a new kind of evaluation culture that emphasizes the efficient production of evaluation data and its active use in formulating and monitoring educational policy. It is important in such a situation that evaluation research recognizes and accepts its own responsibility for the consequences of educational evaluation. (Nowakowski, 1990; Scriven, 1994.) How, then, does national evaluation affect the operations and activities particularly of schools, teachers and students? Evaluation is a powerful weapon in the hands of those who use it. We should always be sensitive to how the evaluation is received by those evaluated, how its findings are interpreted in the schools and how the messages that the evaluation conveys affect them.

In Finland schools have, as a rule, taken a rather favorable view of evaluation studies. A school refusing to participate in an assessment has been a fairly rare exception. As researchers we want to preserve this valuable relationship. It can be broken only once, after which researchers would find it difficult to gather evaluation data from schools, because preconceived attitudes toward national evaluation projects easily turn negative unless those running the projects are able to listen to the schools and co-operate with them in preparing and carrying out the projects.

Open Evaluation and Assessment

Educational evaluation and assessment should be *open* activities and their subjects must be able to make themselves heard in all phases of assessment. Altogether, the present situation foreshadows also the ethical dimensions of educational evaluation. (House, 1990; Meadmore, 1995.) In many countries, curriculum reforms carried out in the last few years have been characterized by an emphasis on the rights and responsibilities of schools and their surrounding communities regarding decisions about the content and implementation of teaching. If we stop to consider the development of our society and the demands that it makes on its citizens and leaders, it becomes obvious that there are very solid arguments to justify such a trend. Educational evaluation must not counteract this trend. Accordingly, a central role of the research community should involve critical analysis of such effects. We must not become mere passive onlookers of the development of evaluation systems and implementers of individual projects. As researchers, we are ultimately responsible to the children and adolescents, whose current and future learning will be determined by decisions guided by our actions.

I rather doubt that the findings of national or international evaluation projects could really lend themselves to drawing straightforward conclusions from educational policy of a kind that would enable politicians and educational authorities to make efficacious decisions about developing education. Learning is simply too complex a process for this kind of straightforward maneuvering. This complexity is increased by the fact that learning is rooted in values: There is no single unanimously accepted definition of learning of high quality or of the kind of actions that would represent its practical result.

At their best, evaluation projects generate knowledge to be drawn on in discussions about the condition and outcomes of schooling. If this leads to the emergence of factual interpretations of how comprehensive school has performed its task of promoting both individual growth and social development, then evaluation has done its work properly. Central in this context, apart from the results of evaluation as such, is also what kind of issues, what objectives serve as the focus of our assessment. Here we cannot but encounter once more the fact that we need many different and complementary forms of evaluation. On the international level we must, despite substantial efforts to develop our methods, rest content with assessments based, in one way or another, on a lowest common denominator. It is difficult to fit the broad cultural spectrum of education and training into such projects, and the compromises that must be made may often be, from the point of view of any individual country, rather severe. (Airasian & Gregory, 1997; Stufflebeam & Shinkfield, 1986.)

A Complementary Process

The relationship between national and international evaluation projects may be derived from the above description. They should be consciously designed to be different and complementary. International evaluation projects, especially the OECD's PISA (Programme for International Student Assessment) for producing student achievement indicators on a regular basis, are likely to create a kind of general framework also for the national evaluation of education. The data generated by the program will help us to get a general picture and place national education and its outcomes in an international context. These indicators will also give us valuable knowledge of some characteristic features of our own education, perhaps with its strengths and weaknesses. This knowledge could serve as a starting point for national evaluation studies. It would make sense to focus limited national resources on trying to identify the causes of the weaknesses affecting our national system and on finding ways of moderating those weaknesses.

National education policy should become sensitive to the messages of international indicators and quick to respond to them by launching national precision assessments and developmental projects (Airasian & Gregory, 1997; Lim & Tan, 1999). It is worthwhile to mention science education in Finland as an example. Back in the mid-1980s, the Second International Mathematics and Science Study revealed very clearly the inconsiderable provision of experimental teaching in Finland, which should have alerted us to looking for the reasons behind and for ways of redressing it. (Serious action followed much later.) International indicators also make it possible for us to assess the impact of national decisions about educational policy. For example, the Third International Mathematics and Science Repeat Study launched by the IEA in 1998 provides an interesting opportunity to assess the immediate effects of the substantial investments made in teaching in this field. Similarly, OECD indicators will later allow us to evaluate the long-term impact of such measures.

The importance of close communication and collaboration between the research community and public authorities is highlighted by the new challenges now facing educational evaluation. As regards preparations for and the technical implementation of evaluation projects, the criteria for successful assessment are, for the most essential part, the same as those for good research. However, the most characteristic role of the research community, the role that most adds value to educational evaluation as a whole, is doing in-depth analyses of very carefully defined sectors of teaching and learning. An understanding of, for example, how the teaching practices of schools could be changed requires thorough and many-sided analysis of and reflection on the reality in which today's schools operate.

In the 1990s, it also became obvious that educational evaluation and particularly evaluation research call for an increasingly professional approach (Brown & McCallum, 1997; Scriven, 1994). As regards methodology in particular, the quality standards set for international evaluation projects are more rigorous than ever. On a national level, meeting such standards requires purposeful establishment of professional research environments. There must also be enough continuity in national level activities to ensure that the expertise accumulated in international projects will be effectively exploited on both the national and the international level. This will also create opportunities for substantially improving our ability to exert influence when joint projects between different countries are being planned.

In a small country like Finland, the foreseeable evaluation projects alone presuppose the expertise and active contribution of so many people that we cannot afford to disperse our scarce resources. From the point of view of the IER, the evaluation projects already in view raise the question of researcher training. Where will we find qualified people to carry out all the coming evaluation projects?

References

Abbott, D., Broadfoot, P., Croll, P., Osborn, M., & Pollard, A. (1994). Some sink, some float: National curriculum assessment and accountability. *British Educational Research Journal, 20,* 155-176.

Airasian, P., & Gregory, K. (1997). The Education Reform Act of 1988. *Assessment in Education, 4* (2), 307-314.

Bonnet, G.(1996). Effects of evaluation procedures on educational policy decision in France. *International Journal of Educational Research, 25* (3), 249-256.

Bonnet, G. (1997). Country profile from France: Profiles of education assessment systems world-wide. *Assessment in Education, 4* (2), 295-306.

Brown, M., & McCallum, B. (1997). The validity of national testing at age 11: The teacher's view. *Assessment in Education, 4* (2), 271-293.

Butterfield, S. (1995). Educational objectives and national assessment. Buckingham: Open University Press.

Cuttance, P. (1994). Quality assurance in education system. *Studies in Educational Evaluation, 20* (1), 99-112.

Daugherty, R. (1995). National curriculum assessment: A review of policy 1987-1994. London: Falmer.

Dylan, W. (1996). National curriculum assessment and programmes of study: Validity and impact. *British Educational Research Journal, 22* (1), 129-141.

Eisner, E. (1984). The art of educational evaluation. London: Falmer.

House, E. (1990). Trends in evaluation. *Educational Researcher, 19* (3), 24-27.

Lim, E., & Tan, A. (1999). Educational assessment in Singapore. *Assessment in Education, 6* (3), 391-404.

Meadmore, D. (1995). Linking goals of governmentality with policies of assessment. *Assessment in Education, 2* (1), 9-22.

Nowakowski, J. (1990). Exploring the role of professional standards in evaluations: Areas and needed research. *Studies in Educational Evaluation, 16,* 271-296.

Sizmur, S., & Sainsbury, M. (1997). Criterion referencing and the meaning of National Curriculum assessment. *British Journal of Educational Studies, 45* (2), 123-140.

Scriven, M. (1994). Evaluation as a discipline. *Studies in Educational Evaluation, 20* (1), 147-166.

Stufflebeam, D., & Shinkfield, A. (1986). Systematic evaluation. Boston: Kluwer-Nijhoff.

Williams, J., & Ryan, J. (2000). National testing and the improvement of classroom teaching: Can they coexist? *British Educational Research Journal, 26* (1), 49-73.

Kimmo Leimu
Institute for Educational Research
University of Jyväskylä
Finland

Conclusions, Challenges and Visions

Two principal problem dimensions at this point of developing international assessment activities are especially seen as warranting in-depth discussion: one concerned with *research aims vs. policy indicator interests* in the system evaluation arena, the other with *international vs. national interests* in conducting large-scale evaluation studies. Both of these problem areas have to do with national strategies and implementation arrangements because any co-operative international activity will ultimately and necessarily be based on national resources and work contributions. This involves, not only priorities and strategies, but also *capacity-building* – the recruitment and training of the necessary human resources – and eventually the problems of organization and funding.

What follows is a review of some highlights of the conference presentations and discussions, with an effort to organize them into main areas of attention and concern. Some liberties are taken in order to go beyond these immediate references in trying to capture certain underlying issues considered pertinent.

Contextual Aspects

In review of the situation, certain trends and policy developments seem to emerge. In general, there is a genuine *need to be aware of and account for the gross societal, ideological and educational processes* that are going on in the world and in several nations. Such needs are recognized more openly and clearly now than they were only a decade ago.

The Ideological, Economic and Managerial Context

There is a genuine need to be aware of the gross societal, ideological and educational processes that are constantly going on in the world and in most modern and developing nations. Such trends are important frame factors as regards the relevance and usefulness of evaluation: How to assess the consequences of different cultural and policy trends in schools? How do they affect educational decisions on inputs and processes, and eventually student learning? – This perspective was introduced in the opening address by Lundgren (in the present publication), who referred to modernization and historical developments in evaluation. There is indeed reason to take a comprehensive look at the general situation. Due to its visionary nature, the review by Torsten Husén (1982) still seems pertinent today and may present appropriate early reminders for the benefit of the present discussion. Husén takes up the following *trends* which were visible as early as 1970:

a) An explosive growth in the demand for education. This is evident throughout the world, and particularly in the fields of secondary and higher education. There has been a shift in demand toward lifelong and adult education, and more profoundly toward considering *education as an investment* whose allocations over the life cycle of an individual warrant careful consideration. The long initial, "once and for all" education pattern can hardly be considered a valid solution any more. This issue has relevance to the organization of various types of education, as well as their consequences for qualifications, career structures and salary principles.

b) Bureaucratization. The growth in educational systems has resulted in a corresponding – and sometimes proportionally larger – growth in administrative structures. There are simultaneous tendencies of growing size in institutional units (schools). This may increase efficiency, but at the same time opportunities for personal contacts are diminished, resulting in the need to increase counteractive measures, such as formal control and supervision.

c) Educational technology and teachers. Developments in the use of educational technologies have been perceived as one response to overcrowded schools. Husén, however, is explicit about the overwhelming importance of human interaction in the teaching/learning process, in which nothing can replace the human element. Such influences are both direct and indirect: It is the task of teachers to provide the necessary learning opportunities to students and to guide their development in an emotionally favorable climate. Herein lies the teacher's professional and ethical responsibility.

d) Meritocratic tendencies. Since the 1960s, education (together with research) has been regarded as the principal determinant of national economic growth, as well as an individual's career development. This has resulted in increasing pressure on the *quality of the final products* of education – and thus on learning achievements. The pressure is felt at ever younger age levels. This meritocratic atmosphere has *gained strength* in the years after the original (1970) predictions. One of its consequences is that studies are motivated by external rewards and other pragmatic goals, rather than regarded as having value in themselves. Husén considers this instrumental career orientation and pursuit of efficiency a price we obviously have to pay for economic growth and material welfare.

e) Research and development are becoming a "knowledge industry". This emphasizes the reality and consequences of the "knowledge explosion", whereby the volume of scientific publications is estimated to be doubling at 5-7 year intervals. It is evident that skills related to the *creation and processing of knowledge* (rather than the absorbing of knowledge) will have increased importance. This leads into increasing uncertainty in mastering this abstract and intellectual environment, with all its practical consequences.

f) Specialized knowledge, while increasing rapidly, also becomes rapidly obsolete. This situation further highlights the need for certain basic knowledge and skills to be gained in formal education. It increases the importance of *independent learning skills and critical thinking* related to problem-solving needs. Flexibility and new adaptive applications are ever more important, as is the notion of recurrent and lifelong education.

g) The responsibilities of schools and formal education are widening – but in competition. With changes in the structure and role of the family, school has come to be considered *the principal agent of socialization and learning*. However, it is *in competition with* other socializing agents, notably the world of the media. Thus, personality development objectives in school education need increasing attention. Husén claims, however, that there are signs of a renaissance of the family, which would necessitate a rather emphatic consideration of the roles and responsibilities of the various partners in education. This has implications for both school and out-of-school learning.

h) Extended voluntary participation in education by the youth. This will have its effects particularly on secondary level education, where *career decisions are postponed*, creating pressure on formal institutions. Key problems here are:

- assuming personal responsibility for one's own education;
- the relevance of the content of education and training to the labor market and working life.

Husén claims that – especially in secondary education – *an "educational disaster area" is emerging*, whereby those willing to face the challenges of the meritocratic society by far outnumber others, who then become an "educational underclass". – In fact, such fears are perhaps more evident today than at the time they were first expressed. Whether this is another necessary consequence of welfare development or – even more seriously – of the dominant values, the validity of some noble goals of previous years may have to be reassessed. To put it simply, this may involve decisions concerning *quality versus equality.*

In addition to visions such as Husén's above, the following observations have become evident:

i) Nations are not alone. Internationalism is not only a slogan and an ideal, it is part of our present and future reality – often very harsh and competitive. Students and the workforce are crossing borders, not only as tourists and cultural ambassadors, but as serious partners, workers and new citizens. We must therefore ask whether children are learning the right things (the content and skills) in our schools. While there are certain national priorities and idiosyncrasies, societal goal-setting and resources are nowhere unlimited, and certain common values, goals and standards are to be observed. Among these are several important feasibility and value considerations that pertain to education: economy, efficiency, effectiveness, quality, and equality... This set of values is one of the points where we are looking into the future, trying to see and predict where society and the world are going.

j) Increasing international competition. Internationalism not only introduces freedoms and possibilities – it also brings about new challenges and straightforward demands, because the extended environment also tends to be competitive. According to Lehmann (in the present publication), the focus of attention in system evaluation has undergone a noticeable change lately, with subject-matter achievement and subject-transcending competencies once again given high priority. One of the reasons is associated with the growing awareness of *global competition*, with a strong view of *education as a source of productivity*. This tends to bring along a quest for *commonalities* and *equivalencies* across borders. Such requirements may concern quality in general, but will soon come down to issues of a common content and common skills, requirements set as well as competencies required. In so doing, all this will necessarily touch upon issues of prevailing values, resources and pedagogies hoped to bring about wholesome development in individuals and entire nations.

k) National concerns for education. A national and local basis of educational development; expectations for international studies. However international any educational needs or assessment results might ever be, their consequences will necessarily be *rooted in national contexts and realities*. Every country has its own history, its values and ideologies, and its own visions of the nation's future. Accordingly, in every country, certain questions of current interest gain prominence and require their particular answers. The most meaningful reflection on and evaluation of any international results will eventually take place in national contexts, where the fundamental rationale and history is better understood. Quite especially, it can be claimed that essential *decisions in response* to observed results (especially shortcomings) are usually possible *only* at the national and local levels. After all, *the significance of evaluation is not in its results, but in its consequences* – i.e. its meaningfulness and role in steering subsequent action!

Accordingly, closely linked with the basic quality and accountability demands is the more research-oriented need to *observe and understand the national cultural and policy context* – i.e. the common values, interests and historical processes underlying any comparative findings – in order to *interpret and learn from* them. While a variety of countries have become involved in cross-national studies, rather different purposes and development trends come in the picture and warrant serious consideration. Notably, the tendencies of either *increasing or restricting local freedom and responsibilities,* mentioned above, will affect directly or indirectly the social and educational realities that need to be understood. Such variety among processes and their potential effects tends to make programs and outcomes of system evaluation intrinsically interesting, but also more demanding.

l) Increasing demand for resources and accountability in education. Education is dear to most nations – both symbolically and economically. Therefore, also the proper utilization of funds has become a *legitimate concern* of those responsible for the planning, management and implementation of educational programs. However, this is not enough, since due to the pervasive nature of education in modern societies, more detailed and reliable *knowledge of the effort and its outcomes* is seen as an indispensable part of the tools needed for development, but also for the maintenance of credibility, authority and power. This is necessary for not only those who are accountable for decisions and actions in education, but increasingly for anyone involved in and affected by the process.

Secondly, *budgetary constraints* on educational expenditure are making themselves felt and highlight the notions of added value in education and of accountability for spending public funds. The very term "education" is hereby being redefined: *Increasingly*, it is perceived as a right to a public service, *decreasingly* it refers to an obligation of the individual to fulfill formal requirements as set by the state (the traditional version) or the obligation of the individual to perfect himself or herself (the 'enlightened' version). This leads to the pivotal role of the concept of *accountability*. In this context, system evaluations appear not only 'justified'; they are *inevitable* if the stability and the legitimacy of the social and political system itself are to be maintained. (Lehmann, in the present publication.)

m) Increasing local functions and freedom. In several countries, society has gone through a considerable *decentralization process* resulting in a rather new and different administrative system, where the number of norms and regulations has decreased. At the system level, the changes are few although fundamental. But at the school level, there are many contradictory developmental processes going on simultaneously. For instance, in Finland schools are expected to be at the same time transparent and open to society but also flexible and innovative, with students having more opportunities to make their own choices. While *ownership* is a very important consideration, schools have the right (they are even expected) to establish their own profiles, but at the same time they must produce relevant common competencies and economize, saving public resources.

n) Quality concerns/demands. Changes in welfare societies, in which the effectiveness of public goods is emphasized, create *new demands on education* and especially on its outcomes. It is in this context that *new demands on estimates of quality* have been made. (Lundgren, in the present publication.) Increasing freedom has lead to greater inventiveness at the local level. One consequence of this greater variety is that some proof of *acceptable quality* becomes necessary. Some units or persons are usually considered accountable (i.e. responsible) for achieving (either facilitating or neglecting) the mainstream goals of learning, but also other virtues such as fairness and human rights. Such considerations are not only necessary in individual cases, or in specialized sectors of education, but extend to the more general provisions of education for all. On this account alone, both national and international comparisons are arousing increasing interest and gaining in significance in a number of countries – i.e. they are *becoming a recognized element* in monitoring the attainment of not only educational, but more generally societal goals, including economic competitiveness and cultural offerings. By providing an international context, one might say, *the whole world can provide a comparative reference group to national assessments!*

p) The institutionalization and regularization of system monitoring. Whereas earlier comparative (IEA) research was largely based on private, individual or institutional initiative and was open to the good will and generosity of various funding agencies, *governments today seem to be prepared* to take serious and active responsibility for monitoring their own efforts. This means not only increased recognition of and interest in systemic educational evaluation in general, but also more stable and more adequate funding for the work itself. The most full-fledged outcome of this devotion is the emergence of the OECD educational indicators project INES, and especially its Programme for International Student Assessment (PISA). Such developments may be regarded as a cure for some of the difficulties of the past: those of a cumbersome start and slow delivery, when funding is thin and uncertain. This widespread formal interest and initiative also holds promise of more serious attention to be paid to dissemination and of our possibility to cope with the consequences of research findings. On the other hand, it may involve certain threats, which will be referred to later in this chapter.

National Characteristics of and Concerns for Education

Any educational system is something that is very *national*. It is also a *political system*, based on policy decisions. Thus, political decision makers have the need and right to have as valid information as possible on the situation, e.g. what schools are doing with taxpayers' money, and whether they are doing their best. These constitute important frame factors with regard to the relevance of evaluation. Therefore, it is a special duty of the national evaluation system to give valid information on what is going on in schools and what the outcomes are. However, the logic of an educational system differs from that of industrial production and from that of other social services. Being a human institution, school always deals with human aspirations and relations, which makes it difficult for us to evaluate how the system is working. (Yrjölä, in the present publication.) Unless national level insight is shown, both the context and the quality criteria of local efforts may remain lacking and their interpretations 'imaginary'. Various types of information (local, national, cross-national) should therefore be made available in a coordinated manner with a view to providing the necessary basis for conclusions.

The Educational Context

The General Idea of Education

According to Lundgren (in the present publication), in earlier days the idea of schooling was to *reproduce* a lost world, *not to produce* the future. This way of thinking has extended even to modern times: Curriculum design is still mainly thought to express the *content* in the curriculum, spell out the syllabus, and to provide guidelines for producing curriculum material. However, changes in knowledge structures, the rapid growth of knowledge – not least improved *access to* information and knowledge – have changed also the notion of curriculum design. Instead of *training* for relatively well-known roles and tasks in stable environments, the new idea about *education* becomes a construct telling about education that prepares students for further needs and forms of education and learning for the unplanned – even the unknown. A main challenge is how to accomplish this, and at the same time legitimize the adopted goals and content in pragmatic terms.

Quality and Productivity Concerns

In Lehmann's terms (in the present publication), the notion of *educational productivity* can be expected to be close to the center of concern in present-day activities in the field of system evaluation. It focuses on the relationships between the inputs and outputs of the system and tries to model the mediating processes in order to facilitate interventions which maximize the outputs under given external conditions. Theoretically, this mode of thinking converges with recent fields of study, such as research on *educational effectiveness* (cf. Scheerens & Bosker, 1997), new models of *educational management* based on economic theory (cf. Dubs, 1996) and *theories of innovation* (cf. Fullan, 1991). In practice, it is highly compatible with the more general concept of *accountability* in democratic societies. It is perhaps this notion which is particularly important in that it transcends the dimension of more technocratic system management in the direction of socially responsible decision-making. We are now moving toward a new kind of evaluation culture that emphasizes efficient production of evaluation data and its active use in formulating and monitoring educational policy.

New Content and New Methods

There is a growing interest in subject-matter learning and learning achievements (Lehmann, in the present publication). However, since the domain of cross-curricular competencies and skills for life in general are also getting attention, one might say that the entire concept of the content of education is experiencing substantial expansion and reformulation, thus presenting ever new challenges to systematic and comparable measurement.

Particularly, important new challenges to both national and international evaluation are presented by the role of the new technologies in opening up the educational system and learning environments. This is not merely a question of new resources, materials, or methods, but of opening up the entire context of education, its learning environments and pedagogical culture. Today, schools co-operate across borders, and children already work on joint projects with their peers abroad. Such network-supported learning environments force us to broaden our views on learning to cover more than just the traditional notion of subject-specific skills and knowledge.

We should also venture into assessing learning that takes place outside of school. Consistently with the notion of lifelong learning, attention has been paid to the demands arising from adult and working life, in terms of choosing the targets of evaluations, but also in defining school learning. While the plausibility of transfer effects has been fading away, the notion of authenticity and applicable skills have gained in significance. The key competencies required in the near future for work, active citizenship and continuous learning appear very similar, be they defined nationally or internationally. (Linnakylä, in the present publication.)

Conceptual Issues

The Evaluation Concept

There is some uncertainty about the content and use of some important terminology, including the concept of evaluation itself. While in Finland *evaluation* is used as a generic term for monitoring and assessment purposes and procedures at different levels, its specialized connotations in the Anglo-Saxon world are somewhat different. In the U.S., the term *evaluation* is used in a more serious and comprehensive sense, perhaps typically associated with assessing several aspects of entire programs or entities, and providing serious statements of their successes and failures. The term *assessment* is not quite as ambitious, and can therefore be more easily used for a variety of feedback information, including student learning outcomes.

The role of assessment and evaluation in education has long been recognized. However, the main concerns have traditionally been at the fundamental level of teaching and learning, where the focus is on individuals and their credentials. Accordingly, student assessments and examinations have long dominated the assessment scene, not least as instruments of selection and rejection. This "case-oriented" (idiographic) and control heritage is still visible, especially as regards the ranking and labeling of particular persons, institutions, or countries, without asking for diagnosis, elaborations, or deeper understanding. Such tendencies are indications of a simplistic conception of evaluation as a tool of authority and power, rather than as one of professional diagnosis and well-reasoned support. (Cf. Norris, 1995.) Its further consequences often remain neglected, and thus remain outside of immediate concerns. (Pettersson & Wallin, 1995; Vedung, 1995.)

Interest in the *evaluation of programs or entire educational systems* is of fairly recent origin – by and large a post-World-War-II phenomenon, when major structural system developments started to take place. It may therefore be considered one aspect of *modernization* in education – or more generally of a new culture. Although one might accept system evaluation as being among the youngest fields of specialized expertise in any nation's educational endeavor, it is, however, no more virgin territory either in terms of preaching or practice.

Evaluation Has Its Contextual Determinants

Like education itself, *the concept of evaluation can be viewed in context and as part of modernization trends.* As such, it is an element in the development of society and of modern thinking. Evaluation is a practical way in which knowledge is extracted from practice. At the same time evaluation is a prerequisite for making reasoned choices. (Lehmann,in the present publication; Lundgren, in the present publication.) While issues as complex as those of education can hardly be described in simple terms, the mere acceptance of a serious use of feedback information necessitates a particular environment: *a climate open to change and rational reason-*

ing. While respecting local and individual integrity, certain *democratic principles* of transparency, openness and valuation of the common good are esteemed. Taken further, this welcomes an *evaluation culture in which a wide variety of interests and paradigms are accepted and nurtured*, including both systematic and representative data ideals and less formalistic participatory approaches. Practical imperatives – perhaps further aspects of modernization – may add other criteria to this, such as productivity, effectiveness, or economy.

In the context of *structural reforms*, the mainstream evaluation issues led to a strategy in which two alternatives were usually compared. In so doing, evaluation research was thought to deliver answers based on comparisons – notably statistical comparisons under non-experimental conditions. This formed a tradition in which comparisons seemed possible irrespective of different circumstances. During the seventies, several new models for evaluation were born. There were rather extensive methodological discussions about quantitative and qualitative methods, among them the case study methodology. (Lundgren, in the present publication.)

When the main structural system reforms were over, the questions addressed changed character. Now *the management and processes of education* were brought into focus. How to construct goals and how to monitor and evaluate the fulfilling of those goals became essential questions to evaluators. Aside from such issues (of implementation), others dealing with contingencies (or causal relationships) emerged. (Cf. Stake, 1967.) This development may be seen as the beginning of (conclusion-oriented) evaluation research. (Cf. Cronbach & Suppes, 1969.) It has its roots and ambitions in *the pursuit of learning from* various local and national initiatives and experiments.

The fourth generation conception of evaluation (Guba & Lincoln, 1989, referred to by Linnakylä, in the present publication) endorses the constructivistic and socio-cultural approach, where different individual, national and cross-national values and principles are negotiated with a view to forming a shared framework for evaluation. This notion paves the way for a systematic view and cross-national studies. A special emphasis is laid on the rights and protection of the target (object) of evaluation (Norris, 1995). Cultural pluralism is recognized in its eclectic and interdisciplinary forms. Evaluators are expected to have content knowledge accompanied with expertise in evaluation theories, methods and practices as well as psychological and societal understanding. Multiculturalism encompasses multiple methodologies, including both quantitative and qualitative approaches and integration of different theoretical frameworks and evaluation procedures (Guba & Lincoln, 1989; House, 1990; Scriven, 1994).

These ideals may be seen as increasingly pertinent at the present stage of modernization. While such efforts warrant due professional attention, the quality of investments and outcomes should also be provided with some *recognized proof* of their effects – e.g. that provided by rigorous and systematic evaluation studies. This requirement would not necessarily imply external evaluation or public sanctions, but might simply involve means to *"unearth" important innovative practices and actual experiences,* which should not be left unrecognized, nor the knowledge dormant. It would seem necessary to bring about increasing professionalism in educational evaluation, not only as far as technical terms are concerned, but more profoundly in terms of its basic values and purposes when developing an entire *evaluation culture*.

The Indicator Concept

The educational indicators of the OECD yield information at the system level about the economy, processes and achievements of education as well as about attitudes toward education. For each of these components, the aim has been to produce a number of indicators reflecting essential aspects pertinent to schooling and education. We in education might wish to figure out how we could have some of the power that gross national indices in economics, such as gross domestic products or unemployment rates, have in the images that policy makers and the general public have of society. How could we capture some of that for education? The significance of the indicators for educational policy, the meaningfulness and attraction of the pieces of information conveyed, however, *depend ultimately on the needs and views of the user.* The OECD indicators are system level indicators. There are needs for a variety of indicators at other levels as well. (Owen, in the present publication.)

We always tend to have the difficulty of different levels and kinds of indicators, and also who the indicators are intended for. One of the thorniest issues is how many indicators we really need. For example, we might wish to have what some would call *social indicators*, some general level of income and an idea of size, and so forth. Or, we might have what we call *participation indicators,* what we actually classify as *inputs*: how many children go to school, enrollment at the different levels of education, etc. Likewise, we have something that belongs to *processes and institutions.* Much of this would constitute the domain of contextual variables and indicators. Finally, we have indicators of *student achievement* and *attainment,* which is looking at what kind of graduation rates we have, what kind of school completion rates we have at different levels, and how we are achieving in, for instance, mathematics? However, people are more and more interested in yet another kind of student outcomes. We call these *"cross-curricular competencies"* – those goals and skills which are not subject-bound, but help to lead to success in later life. These might not be measured as reading, mathematics, or science. (Owen, in the present publication.)

There are a number of new challenges presented to the future work on indicators. These include authentic tasks and skills developed and useful outside of school, notably those of collaboration and distributed knowledge. Even the notion of cross-curricular competencies may need to be reviewed from a lifelong perspective. The problem more generally is that for many of the outcomes we just do not have an indicator system yet. Still, we have to be able to measure them, in order to have an indicator. It is not likely that we can develop all the conceivable measures. Involved is the whole problem of self-reporting of process, of really getting at teaching and learning, and in many ways when talking about education, this is the heart of the matter. What teaching and learning are like is really the black box in all of the indicator programs. This is a continuing challenge, a continuing problem. We are trying to work on the cross-curricular competencies – such as problem-solving and working in groups, using technology – competencies that you do not necessarily learn in any one course but that are part of our idea of production rather than reproduction of knowledge. (Owen, in the present publication.)

International Assessment Studies

A Multitude of Expectations for International Studies on Education

What do the international programs give us? The pioneering IEA projects recognize *two purposes* as far as international comparative achievement studies are concerned:

1. To provide policy makers and educational practitioners with *information about the quality of* their *education* in relation to relevant reference groups, such as the key trading partners. This gives an *overall standing*, which makes certain *comparisons* possible. In line with these general purposes, the IEA strives in its

studies for *two kinds of comparisons.* The first consists of *direct international comparisons* of effects of education in terms of scores or subscores in international tests. The second kind of comparison concerns *how well a country's intended curriculum is implemented* in schools *and achieved* by students. (Plomp, in the present publication.)

In addition, the studies provide *benchmarks*: some notion of the distribution of the population across levels of performance, some kind of description of performance characteristics and of population distributions across them. Further analyses may give us some notion of how particular subgroups – such as *national minorities* – perform compared to certain overall benchmarks. With repeated cycles of studies, we can get *trend data.*

2. To assist in *understanding the reasons* for observed differences between educational systems. This potential is linked with the vast array of contextual background variables and data, which may be used in terms of explanatory models and further theoretical work. (Plomp, in the present publication.)

Reference has been made (by Husén, 1982) to *the world as an educational laboratory* because countries plan and implement their education differently. No doubt, that can be seen happening no matter from which of our perspectives we are looking at it. As there are several possible perspectives and approaches, *the question of audiences* emerges again. Some of these interests, each of which has its own expectations and quality criteria, have been summarized by Leimu (1992) and include:

Policy interests

- Cultural interests: the status and role of formal and non-formal education, modernization.
- International comparisons: benchmarking, strengths, and weaknesses.
- Historical interests and timing: the stages of development of the educational system, changing contexts, extending into the future.
- Accountability: the degree and quality of implementation.
- Economy: the efficiency and economy of education.
- Policy-making: societal phenomena (equity, sub-groups, needs of working life), predictions.
- Administrative-managerial interests: planning, implementation, and leadership in multilevel systems.
- Policy consequences relying upon research studies and indicators?

Scientific and conceptual interests

- Theoretical and construct validity issues: causal postulates and observations.
- The structural perspective: explanatory potentials for the pursuit of quality learning as a complex multilevel phenomenon.
- The curriculum perspective: structural, input and processual prerequisites and guidelines for school learning; intentions vs. implementation actualities.
- Educational-psychological processes: conceptions of learning as a process and of learners as human beings; interpersonal relations and learning experiences; attitude formation.
- Research method interests: What paradigms are effective in descriptions and explanations? How to ensure comparability?

Technical and management interests

- Resource considerations: How to obtain useful, timely and high quality feedback information most economically and efficiently? How to provide for comparative information needs and national refinements?
- Management, supervision, and formal accountability interests: How to plan, organize and manage large-scale research? How is the data organized? What effective use to make, and how, of the results?
- Dissemination interests: What are the target audiences and strategies of making the research results known? How are the results most efficiently put into use?

Inevitably, it may be impossible for any one research venture to prepare us for and respond to all of these needs. However, with properly organized data, a variety of secondary analyses will become possible beyond the initial and core results. Whatever the particular interest, one may suggest that international studies help us *to better comprehend the functioning of our own system* – its specifics, and perhaps its idiosyncrasies. Perhaps very little progress has been made along those lines. There are interesting issues pertaining to *functional equivalencies* between systems and the dynamics of two respective systems in comparison with each other. For instance, one system may apply certain differentiating mechanisms – e.g. centralized matriculation exams – whereas another system applies some other strategies – e.g. competition between secondary schools. – Such comparisons might well be a major challenge in future IEA-type work. (Lehmann, in the present publication.)

In general, research results and international indicators can give people a handle on *what is going on in education in the country and also raise their interest level*. The more accessible the indicator set is and the better we are communicating it, the more likely it is that important issues are taken up and *become part of the national debate*. Indicators, therefore, may be seen as stimuli to and openings of discussion. They do not provide final explanations, but they can point out problem issues and provide evidence for certain matters which can become serious considerations when developing national policies or instructional approaches. In a sensible strategy, they should also point at further in-depth research needs.

Conceptual and Content Comparability

Comparability begins at the conceptual level: agreeing on the common usage of terms and constructs. These are very basic, as one has to reach an agreement in determining certain key concepts, such as: What is a school? Who is a teacher? Who is a student? or arrive at common target population definitions. Typically, this extends to common technical standards, which concern decisions and practices such as sampling, field testing operations, or data management. All these very cornerstone notions are quite important and may often be rather interpretative.

Both the IEA and OECD assessment frameworks and instruments have been criticized for cultural bias or even cultural imperialism. However, the criticism has been strongest among people who have *not* been involved in these efforts, being thus unaware of the multicultural discussions which always take place when defining the basic concepts, when constructing the evaluation framework, and when designing the instruments. Admittedly, the social, cultural, or educational characteristics of the developing countries have not been taken sufficiently into account in designing some of the studies. (Linnakylä, in the present publication.) On the other hand, one might point out that one of the principles necessarily adopted in the IEA work has been the attempt *to be equally unfair to everybody*. (Leimu, in the present publication.) Within the OECD, however, the range of cultural and developmental differences among the member countries is not quite as large, making comparisons between those countries fairer. Inevitably, certain compromises need to be accepted if we want to do this kind of work in a cross-cultural setting.

International evaluation projects are, at their best, *joint ventures*: planned, implemented, and reported together. This has not always been understood in the national context. Instead, it has been assumed that the conceptual-theoretical framework and the instruments come from somewhere as given, and the national contribution is restricted to faithfully carrying out the instructions received. Such a contribution will not, of course, feel very inspiring or challenging. (Linnakylä, in the present publication.) Not only this; it is, indeed, necessary to point out cultural peculiarities early on, in order for them to become properly observed in the design and incorporated in the final measures. Otherwise the results and their interpretations may become incomprehensible or simply impossible. An important remedy would be active, full-fledged participation by all country representatives in the projects right from the beginning, as well as careful national criticism in all phases of instrument development. The process has its natural sequences, ending with open and joint consideration of the results, their contextualized interpretations, and their various implications for policy, theory and practice.

Since instructional processes are among the most sensitive issues and difficult to capture in large-scale research programs, they warrant special attention. In trying to make some progress in measurement, one possibility would be to look at opportunities for moving instructional activities into assessment procedures. This could be accomplished e.g. by creating authentic tasks, favoring performance assessments, and looking at portfolio assessments.

Consequences of Evaluation and Indicator Work

What Can Be Done With the Results?

How, then, do national evaluations *affect* the operations and activities particularly of schools, teachers and students? How are they made available and disseminated? What actions and with what delays can be expected? How are the effects brought about? What is the interplay between system level, local and school actors?

Since every practice and principle has both its merits and drawbacks – two sides to the coin – every experience and serious piece of information, be it local or national, should be set against its intentions and actualities, and also its costs and effects. Since education by its very nature is based on human skills and interactions, these contrasts are natural and always there. In this case, the resources and expenditures in question are both human and material. The compromises to be made are therefore often between what may be deemed desirable and what is actually possible, and one needs to ask perhaps more seriously than ever: *At what human costs* are various practices adopted or outcomes attained?

Knowledge Is Power

It is a well-known fact that *evaluation is a powerful weapon* in the hands of those who are using it. First, we should be sensitive to how evaluation procedures and results are received by those evaluated, e.g. how the findings are interpreted in the schools, and how the messages received from the evaluation come to affect them. (Välijärvi, in the present publication.) Another important agent is the one having high-quality knowledge of and insight on the state (process and progress) of education, the one willing and able to exert active influence on the basis of such information. This is likely to take place in terms of public discourse concerning education. As noted by Pettersson and Wallin (1995), reliable and pertinent knowledge will possess power in public fora, and those in possession of such knowledge may come to exercise publicly recognized authority through their leadership position in the discourse.

Important in this connection will be *the question of consequences,* especially from the viewpoint of those whom the assessment concerns. The issue here is: What kind of responsibilities will the research community have for influencing the *consequences* of national evaluations, especially in terms of their pedagogical and human implications? There is the question of how critical a role researchers are willing and able to assume regarding the evaluation projects to be implemented, and what kind of responsibility they will accept for the conclusions drawn. (Välijärvi, in the present publication.)

It seems natural that the main consequences of assessments are related to the original user interests. Thus for instance, national level assessments should draw the attention of and be used by national policy authorities for revisions of policies, structural characteristics and resource allocations. Likewise at the local level, curriculum-related developments and their financial consequences should be considered. At the individual level, the traditions of assessment are long. However, it would be in line with modern educational thinking that the consequences of positive and supportive action should receive major emphasis, instead of the traditional labeling and other negative practices.

General Requirements for Quality

The Evaluation Model

Educational *indicators* are *necessarily few and selected*. For instance, the early ambitions of the OECD/INES were targeted at 54 indicators, of which half a dozen were assigned to the domain of student learning outcomes. The expected benefit of such a compact approach is that the database can be obtained and analyzed efficiently. This allows relatively frequent and regular cycles. Its shortcomings are in its restricted scope and focus. It is self-evident that in this approach indicators must be carefully chosen and well justified. It may be claimed that this process will remain haphazard and erratic unless there are a well-conceived conceptual framework and empirical experience underlying the total set, giving it a transparent structure and credibility. An overly general model will not be sufficient here, but one that is progressively specified at a meaningful latent variable level (cf. the "information pyramid" notion). It should be noted that this requirement is not restricted to the overall research or indicator model – it applies to each specific domain (content area) included in the exercise. Thus, there is a need for adequate conceptual and modeling work in each of the subject domains, covering not only their structure or their "traditional" and "modern" substance, but also their common pedagogical concerns, such as formal learning objectives. Similarly, a well-reasoned conceptual structure is a prerequisite for any background information and instrumentation, because it is here that the problems of achieving cross-national ("culture-free") equivalence are even more challenging and cures necessary. Since all other indicators may be regarded as constituting the necessary system context, the role of background characteristics specific to school learning outcomes has so far remained restricted. However, there are signs of increasing ambitions here, which tends to make the above requirement even more serious.

On the other hand, *the research approach* (especially of the large-scale survey type) typically aims at relatively *comprehensive conceptual models* and large collections of manifest variables in testing such models. Due to the complex nature of the phenomena under study, multilevel conceptual models tend to be commonplace; the scope of measurement is usually more extensive and innovative; and, hence, in the analyses there is more room for "fishing" and "hunting", i.e. certain explorations in unknown territory. In education, a field which is constantly in motion, such scope and flexibility may be regarded as necessary, if one is to keep in controlled touch with realities. A benefit of this (research) approach is that it leaves several options open and enables constant renewal. Its shortcomings are the slowness of delivery and necessarily incomplete analysis of the abundance of the data available. Thus, the overall effort tends to remain somewhat inefficient – especially if the funding does not allow time for extended and simultaneous work on different fronts.

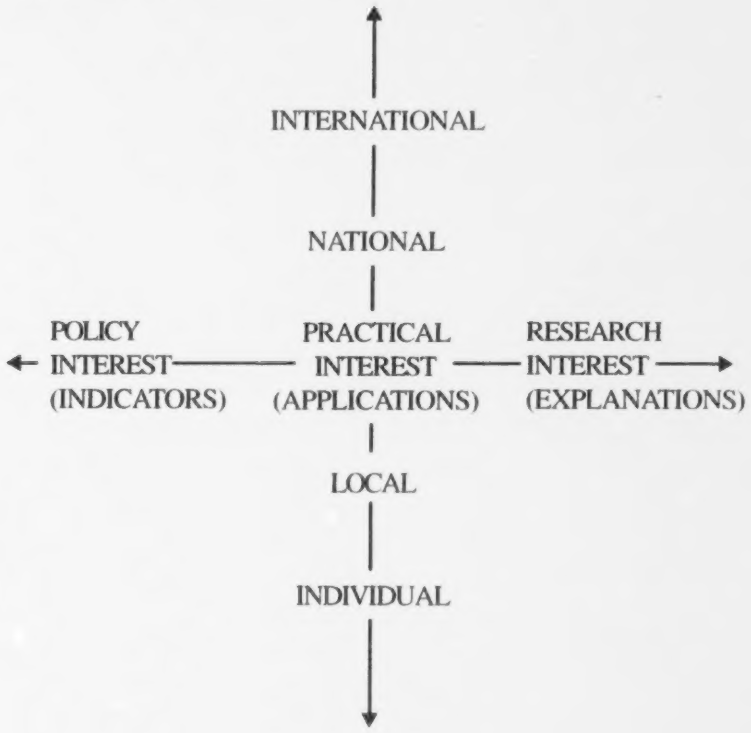
The pioneering international contributor in the field so far – the IEA – was founded as a research co-operative. Therefore, the interests in international comparative studies initially had a research perspective, and curriculum-related modeling was one of its core elements. More theoretical studies have often been carried out at the national level of further analyses. In the second half of the 1980s, there was increased recognition by the IEA of the interests of policy makers and of educational indicators. IEA databases were made available for OECD achievement indicators, demonstrating the direct usefulness of research-based data in the OECD publications.

The issue can be put in simple terms: If we want to know *why*, we get into both qualitative and quantitative research arenas. This is why indicators and research have to constitute some sort of *complementary systems. Research* is the basic foundation we really need, while *indicators* are in some ways a communication tool, they are not the end you are seeking. *What we need is understanding based on research, but we also need to be able to communicate certain aspects of that research quickly, and that is the role for indicators.* (Owen, in the present publication.)

A particular role in which such research-based indicator information is effective is *that if there are particular myths in a country that have been accepted as educational truths, one can present data to counterbalance them.* So this again implies looking a little deeper, having something to say about the processes of education and being specific about it. This can do a lot in changing the nature of debate, and it is here that indicators can help in forming the debate of where to look, of how to go into more depth. (Owen, in the present publication.)

The initial charts (introduced by Leimu, in the present publication) might now be revised to some extent, taking better into account the different levels of operation:

Figure 1. The principal problem dimensions.


INTERNATIONAL
NATIONAL
POLICY INTEREST (INDICATORS)
PRACTICAL INTEREST (APPLICATIONS)
RESEARCH INTEREST (EXPLANATIONS)
LOCAL
INDIVIDUAL

Evaluation and Indicator Strategies

A strategy should basically lay out the general purpose and arrangements of the evaluation: *what* is being done and *why* (e.g. For whom? For what?). It should ideally stake claims to *how* (e.g. By whom? With what means and resources?) and *when* (i.e. At what time? How often? In what order?). All of this might benefit from an orchestration metaphor in which the orchestra plays a symphony or jazz, with its particular solo and ensemble parts, always relying on sufficient professional skills of the players of the various instruments, each playing their particular notes, but together creating a total harmony.

Questions that we might entertain could well have to do with issues like the following:

1. What are the local/national resolutions, especially in terms of *how* and *with* whom? A question concerning national structuring is: Who are the necessary partners, and what are their mutual relationships? How are all *these necessary partnerships* formed and maintained? Are there established responsibilities, and is there something to be gained by each party?
2. While any comprehensive operation of this magnitude and of high quality standards is necessarily expensive, a real question becomes that of *funding, i.e. investments*: Can a member country really afford intensive work? At what and whose expense is the work being done? Are some important elements of evaluation, research, or indeed educational developments neglected in so doing.
3. The other side of that coin is the question of *benefits*: What can be gained by participating? It is evident that there is inherent public interest in the quality of education, which is concerned with entire age groups and extends from childhood to mature life. It is very likely that enormous savings can be achieved if one knows what is actually happening in the educational system. Not only would this meet accountability expectations, but it would also enable active *steering*: consideration of how things could be done more efficiently or with less human suffering. Further issues are those of gaining awareness of current international trends in education and state-of-the-art research methodology. The overall balance warrants careful consideration.
4. How to secure *renewal* in repeated cycles of studies, while capturing trends using old procedures? One ambition lives on extensive search and critical revisions, the other thrives on continuity. The current answer seems to lie in a kind of modular approach in which elements of one study cycle are incorporated in the next ones. This "bridging" allows making comparative estimates between studies, while leaving room for continuous instrument renewal. – Another consideration is renewal of the research competencies involved. It can be said without hesitation that large-scale international studies provide an exceptionally fertile soil for researcher training. Being complex in nature, they always necessitate national and international teamwork, where opportunities for supporting versatile competence-building arise naturally.

From an international perspective, the work of *the OECD in developing educational indicators* entails increasingly systematic international evaluation projects deliberately linked with educational policies. The timetables and quality standards adopted for implementation presuppose highly professional and long-term organization of assessment. The consequent national arrangements represent a variety of approaches and solutions. In Finland the process is nowadays highly centralized. After the public debate, which is usually drawn-out but essential, the National Board of Education arrives at its final conclusions and prepares a set of action points for the Ministry of Education. It informs the Ministry about the main problem areas and the action needed. The Ministry of Education subsequently makes political decisions and the National Board of Education is given the task of more detailed planning and implementation of the improvements. (Takala, in the present publication.)

The results of evaluation can be *used* in several ways. They may contribute to the development of overall educational provision, in particular when developing curricula, teaching methods and teaching materials – important at the school level. They are also useful when allocating resources and in policy making. However, one of the firm principles of the National Board of Education is that the task of *national evaluation* is not to produce ranking lists, but *to identify problems in the educational system.* Evaluation at the *local and school level*, called self-assessment, has the same logic. A school's assessment of itself is the first step toward improving teaching and education and ensuring quality. It is not an end in itself. The conclusions drawn are considered the most effective way of making the necessary changes, but *schools need a more general framework in order to compare their self-evaluations*, to see where they are and where they should go. (Takala, in the present publication.)

The Strategic Approach: Phasing and the Distribution of Responsibilities

A question linked with the previous one is: *What kind of division of labor and co-operation between the research community and educational authorities* will work best in the rapidly changing field of evaluation? What complementary roles could educational research and indicator work play?

It is in *ensuring a high standard of educational evaluation* that the research community has a particularly important contribution to make (Välijärvi, in the present publication). The quality criteria governing educational assessments and evaluation studies more generally have close parallels, particularly as regards the practical implementation of evaluation projects. The shared goal is timely and reliable production of many-sided, useful and generalizable knowledge.

Relationships between *national and international* evaluation projects may be derived from the above description. *They should be consciously designed to be different and complementary.* Due to its currently dominating role, especially the OECD's Programme for International Student Assessment (PISA) for producing student achievement indicators on a regular basis is likely to be in a key position in generating new demands for general frameworks also for national evaluations of education. The comparative data produced by the program should help us to get a general idea of national education outcomes in an international context.

At the same time, national education policies should become sensitive to the messages of international indicators and be quick to respond to them by launching *national precision assessments and developmental projects*. How to do this, is a matter of national level planning according to the existing needs and possibilities. When doing this, it will be wise to work out acceptable functional relationships between the main parties concerned, usually including (national and local) educational authorities, researchers and schools, so that long-term co-operation will become possible. The importance of close communication and collaboration between researchers and public authorities is highlighted by the new challenges now facing educational evaluation. The issue concerning the preconditions for the research community to be able to meet the increasing challenges as regards evaluation itself is another important domain in any general model for assessments, posing also questions on the production of *well-substantiated evaluation data of high quality standards*.

According to Linnakylä (in the present publication), *international and national evaluation schemes should be intertwined* so that they complement each other without straining schools and students too much. *International evaluations* constitute a useful monitoring instrument and one possible basis for orienting national assessments. At their best, international evaluations might include qualitative case studies or action research in order to cast light on quantitative results and draw a nationally sharper and more colorful picture of student achievement and more immediate learning context factors. These substudies should ideally be connected with national developmental interests in instructional processes and education in general.

As for *national assessments*, the primary emphasis should be put on national values, goals and standards, with a view to drawing a more revealing profile of learning outcomes, particularly in the disciplines and content areas not tested internationally. Intensive studies should be focused on those areas where problems are detected in the international findings. (Linnakylä, in the present publication.)

Researchers themselves must define their own role in the field of evaluation and assessment on a new basis. Their strengths are in their long and many-sided experience in implementing demanding research projects that this sustained interest has brought about. The definition and detection of appropriate research partners is not a simple and straightforward matter. There are often several active agents in the field of evaluation and evaluation research, to say nothing of the multitude of interests in different disciplines, or levels and sectors of the educational endeavor. Competition for funds is therefore speeding up, whether or not the volume of funding allocated to evaluation and assessment is increasing. Competition as such should be regarded as a positive matter and is likely to improve rather than narrow the scope of action. However, it is important that the different research communities ask themselves what kind of role they are willing and able to assume in the changing field and challenges of evaluation. (Välijärvi, in the present publication.)

The overall quality of research and indicator data is not only dependent on the work of researchers. In all social research, one comes to deal with human counterparts, not least with students and teachers, who constitute the primary source of information. Their roles and interests are therefore an important consideration in any evaluation strategy. Proper *channels and modes of communication* are indispensable when working with schools in the field. As researchers we must fully comprehend the importance of preserving an *open and trusting working relationship*. In the Finnish case, this is built upon common professional interests and understanding among schools and researchers, and a school refusing to participate in a study has been a fairly rare exception. But the situation may be delicate: The confidence of teachers need only be broken once, and then it will be difficult to restore. Researchers would then find it difficult to gather data from that or any other school, because attitudes toward similar evaluation projects may easily turn negative wholesale. Unless those running the projects are able to properly communicate and co-operate with the schools in preparing and carrying out the projects, long-term difficulties may arise. (Välijärvi, in the present publication.)

Other bases for ensuring participation are also conceivable. These include legal or administrative rules and potential sanctions. Although less desirable bases than professional approval, these may be deemed as strong and effective means. Such a situation is in fact emerging in Finland where the new educational legislation stipulates that schools and municipalities accept considerable responsibilities for participating in external evaluations as well as for doing their own self-assessments. It is therefore important that such activities are *open and transparent*, and that the participants are able to make themselves heard in all phases of assessment, supporting a sense of *ownership*. Without the threat of negative sanctions, the benefits of systematic information should be well understood by all those concerned, and this goal should be well achievable. Altogether, *this requirement emphasizes serious consideration of the ethical dimensions of educational evaluation.*

Other Considerations

Political vs. Educational vs. Research Logic?

In many countries, curriculum reforms made in the last few years have been characterized by an emphasis on the rights and responsibilities of schools and their surrounding communities regarding decisions about the content and implementation of teaching. If we observe such overall developments in society, there are solid arguments for justifying certain trends, and the consequent implications become more obvious. Educational evaluation should not counteract such trends, but instead seek to honor their (humanistic) consequences. Government-sponsored developments are not without problems, however. Although the idea of the system's responsibility for its own monitoring is reasonable, one may put certain questions:

- Will political logic take over genuine research ideals in the future? In particular;
- Will the focused and reduced indicator model replace the more open and broad research approach, leaving us with a few reliable but perhaps insufficient indicators of educational inputs, processes and products? How could these other needs be catered for?

Accordingly, an important role of the research community involves doing critical analysis of general societal processes and their effects.

There are also issues of quality that have to do with *professional integrity*. The underlying issue is *what role researchers will be given in national system evaluation projects* and strategies? What initiatives are expected, and what independence is assigned to them? The more concrete question is whether researchers should content themselves with merely carrying out evaluation projects designed by others, concentrating mainly on competing for the resources necessary for the practical operation? It is especially in sample-based surveys which regularly monitor the condition and effectiveness of formal education that one may ask the question whether the research community should be merely carrying out pre-specified evaluation tasks commissioned by others, or whether they should actively contribute to the use and development of a system of information, on the basis of their evaluation and assessment contributions. (Välijärvi, in the present publication.) Parallels with the changes observed above in the image of education itself are easy to see. In this process, *national evaluation models of education and school learning* become necessary to researchers: They become a basis for their own identity as researchers. This is in keeping with the ideals of explanation and understanding, which are central in the research-oriented approach. Such soul-seeking is continuously necessary.

These concerns relate, not only to differences in *interests*, but also in *degrees of freedom* among different professional groups that may have their roles and stakes in any system of assessment. Administrative officials within educational systems, acting as civil servants, are heavily bound by political and legal decisions. In short, they have to firmly believe in and push their particular cause.

On the other hand, researchers swear by their methods and conceptual views. Accordingly, they can (and should) have their apparent freedom from consensual constraints, as they can (and should) always question whether the adopted values, ideals, content, and objectives are also adequate. Thus, the fundamental standpoint of a researcher is to think critically of his or her approach and findings. It may be suggested that such differences in the roles of different participants in a national evaluation strategy should not become sources of conflict, but be used to the advantage of the totality. While educational authorities should know and understand policy interests and implications, and while the voices of representatives of the field should be allowed to be heard, conceptual-technical expertise and critical innovative capacity should be expected from researchers. – This leads us to further considerations, such as

Problems of Renewal

How to ensure the constant renewal and development of our *evaluation approaches and methods* at both the national and international levels, to maintain their vitality and relevance over the long-term cycles of repeated data-acquisition and use? An equally demanding challenge involves the constant renewal of the *content* of evaluation. How can assessment be made to encompass, as fully as possible, the spectrum of instructional goals? How can it be done with maximum meaningfulness and comparability, and still be feasible? A system of evaluation that is implemented mechanically and whose content remains unchallenged and unchanged fails to convey messages of renewal and becomes an actual threat to the innovative development of teaching. While research is an essential aspect of educational assessment and evaluation, and if the implementing agencies wish to be dependable agents in the field, they must also assume a critical role vis-à-vis evaluation. The challenge facing evaluators is how successfully they are able to combine these two roles (viz. those of stability and change).

As one possible prerequisite, the international databases should be made available to all participating countries as soon as possible, so that the national centers could do further analyses of and extensive research on the data. Such promises are not always easy to fulfill. Yet, the databases and further analyses would be invaluable both in research and in training new evaluation experts. (Linnakylä, in the present publication.)

The renewal of *the cadre of evaluators* can be a problem needing attention. In many countries, it is hard to find and recruit young, sufficiently qualified evaluation researchers, whose repertoire of knowledge and skills must be extensive. At least in Finland, as a result of more fashionable interests such as qualitative research approaches in education - although valuable as such - methodological competence among young researchers has become one-sided, and their knowledge of the newer statistical methods and computer software necessary for dealing with large data sets is limited. By the same token, their knowledge of evaluation theories and implementation strategies needs strengthening, together with competencies to deal with social, economic, and cultural changes. If the obligations of international evaluations are to be met in the near future, serious attention needs to be paid to the further training of evaluation experts and researchers. (Linnakylä, in the present publication.)

In this work, international co-operation becomes a source of enrichment and support. It would be desirable to include researcher training as one component in national and international co-operative schemes, which provide excellent environments for dynamic "learning by doing". This issue deserves serious attention in the coming years, because large international projects offer repeated, yet unique opportunities to everyone involved. One of the problems involved may be the image of evaluation: Evaluation is not considered real science, whereby corresponding researcher training has not enjoyed very high esteem in university faculties and departments of education. This image could easily be improved. (Linnakylä, in the present publication.)

The Legitimacy of Assessments

There is a possibility that the *legitimacy of assessments* becomes questionable if the same agency that is responsible for the system and program planning and implementation is also doing the evaluation.

Also this coin has several sides:

1. It is possible that a *governmental* evaluation project will be perceived as more legitimate and acceptable than a purely institutional (research) effort. This interpretation would be likely in a relatively strong or recent tradition of centralization.
2. It is possible that a *non-governmental* evaluation project will be perceived as more legitimate and acceptable (e.g. by reason of being more independent and impartial) than a governmental effort. This interpretation would be likely in a relatively strong tradition of decentralization. By the same token, it should be recognized that whether under government auspices or not, major evaluation efforts anywhere are likely to be dependent on substantial government support.
3. Potential sanctions are likely to be regarded differently if the study represents the formal system rather than merely information interests. Educational researchers can hardly be regarded as representatives of the Establishment, so their influence can at best be indirect. This could give their work an aura of neutrality. However, due to their lacking formal authority, their "muscle" to exert influence is weak, and hence their rights to gain access to schools may more easily be questioned. On the other hand, formal representatives of central agencies (in Finland, the Ministry or the National Board), while re-

spected and well received in schools, may elicit responses which are purpose-colored rather than genuine.

A Power Game

Other considerations of the consequences of evaluation may concern visibility, authority and influence. The question may be put whether an actual *power game* will emerge among the various parties involved or excluded as a result of divergent ideals and efforts? *Whose data and interpretations will come to dominate public discourse on education?* Who will be in a position to launch large enough projects to successfully plan and carry out meaningful studies? Can such power contingencies be avoided or reconciled?

Usually, this is tantamount to *the sanctions and benefits to be expected* after evaluation. While it is understandable that external evaluations in particular may arouse feelings of suspicion and fear, the need to be explicit – transparent – about their purposes and consequences becomes an imperative. In order to create a climate and attitudes favorable to assessments, it is always important to inform the client of the basic rationale, including the intended benefits and possible consequences. This approach is necessary not only at the introductory stages, where access to schools and students is negotiated, but even more at the stage of dissemination, when the situation may be more sensitive and open to favorable or unfavorable experiences by the consenting partners. While individual *students* are exposed to regular testing and examination practices, whereby they are likely to experience strong emotions, we tend to consider this perfectly acceptable even if sanctions against them may be quite severe. But also at the system level, one may speak about high-stakes and low-stakes assessments in terms of their consequences for the school evaluated. Although these may be subject to the prevailing administrative and management climate, unnecessary tensions should be avoided by being very careful and competent with the ways research information is made public. The power exerted by enhanced knowledge should not be used as a weapon of unhealthy competition or unnecessary sanctions; rather, it should represent professionally sound judgement and be well justified, appropriately channeled for development purposes.

References

Cronbach, L. J., & Suppes, P. (1969). *Research for tomorrow's schools: Disciplined inquiry for education*. London: Collier-Macmillan.

Dubs, R. (1996). *Schule, Schulentwicklung und new public management.* St. Gallen: Universität St. Gallen, Institiut für Wissenschaftspädagogik.

Fullan, M. (1991). *The new meaning of educational change.* New York: Teachers College Press.

Guba, E. G., & Lincoln, Y. S. (1989). *Fourth generation evaluation.* Newbury Park, CA: Sage Publications.

House, E. R. (1990). Trends in evaluation. *Educational Researcher, 19* (3), 24-27.

Husén, T. (1982). Present trends in education. *Prospects 12* (1), 45-56.

Leimu, K. (1992). Interests and modes in research utilization: The Finnish IEA experience. *Prospects 22* (4), 425-433.

Norris, N. (1995). Koulutusohjelmien arviointi [Evaluation of educational programmes]. In S. Takala (Ed.), *Arviointi ja koulutuksen laadun kehittäminen* [Evaluation and the enhancement of the quality of education] (pp. 33-38). Jyväskylä: University of Jyväskylä, Institute for Educational Research.

Pettersson, S., & Wallin, E. (1995). Utvärderingsmakt [The power of evaluation]. In B. Rombach, & K. Sahlin-Andersson (Eds.), *Från sannigssökande till styrmedel. Moderna utvärderingar i offentlig sektor* [From seeking after the truth to steering methods. Modern evaluations in the public sector] (pp. 93-108). Stockholm: Nerenius & Santerus Förlag.

Scheerens, J., & Bosker, R. J. (1997). *The foundations of educational effectiveness.* Oxford: Pergamon Press.

Scriven, M. (1994). Evaluation as a discipline. *Studies in Educational Evaluation, 20* (1), 147-166.

Stake, R. E. (1967). The countenance of educational evaluation. *Teachers College Record 68* (7), 523-540.

Vedung, E. (1995). Utvärdering och de sex användningarna [Evaluation and its six uses]. In B. Rombach, & K. Sahlin-Andersson (Eds.), *Från sannigssökande till styrmedel. Moderna utvärderingar i offentlig sektor* [From seeking after the truth to steering methods. Modern evaluations in the public sector] (pp. 25-51). Stockholm: Nerenius & Santerus Förlag.

Merging national and international interests in educational system evaluation

This publication is a child of love of one rather specialized sector in education - that of acquiring and using empirical information as a basis for monitoring and studying education with the special ambition of making such information both meaningful and powerful, and its use dynamic. This pursuit has a distinguished past, an intriguing present, and a challenging future. In particular, the present publication has been inspired by two major approaches to these ideals:

The work of the IEA, the International Association for the Evaluation of Educational Achievement on comparative educational research, which started in the 1950s and is going on.

The other and more recent one is the OECD/CERI project INES for developing and acquiring Indicators of National Education Systems, especially its strategy Network A, which aims at collecting student learning outcomes data on a regular basis.

We start discussing educational system assessment from various perspectives - both national and international, and as regards both educational research and policy. This is no more a completely open field, but one in which certain premises and traditions have already been established. At the same time, several issues, problems and challenges are still waiting for further discussion and solutions.

INSTITUTE FOR
EDUCATIONAL RESEARCH
UNIVERSITY OF JYVÄSKYLÄ

BEST COPY AVAILABLE

ISBN 951-39-0915-8

Cover: Martti Minkkinen